

Technical Report 1174

**Army Enlisted Personnel Competency Assessment
Program: Phase II Report**

Deirdre J. Knapp and Roy C. Campbell (Editors)
Human Resources Research Organization

20060313 011

January 2006



**United States Army Research Institute
for the Behavioral and Social Sciences**

Approved for public release; distribution is unlimited.

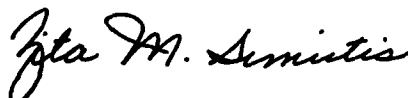
**U.S. Army Research Institute
for the Behavioral and Social Sciences**

**A Directorate of the Department of the Army
Deputy Chief of Staff, G1**

Authorized and approved for distribution:



**MICHELLE SAMS
Technical Director**



**ZITA M. SIMUTIS
Director**

Technical review by

Kimberly Owens U.S. Army Research Institute
Peter M. Greenston, U.S. Army Research Institute

NOTICES

DISTRIBUTION: Primary distribution of this Technical Report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, Attn: DAPE-ARI-MS, 2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

FINAL DISPOSITION: This Technical Report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this Technical Report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

1. REPORT DATE (dd-mm-yy) January 2006			2. REPORT TYPE Final		3. DATES COVERED (from... to) November 2003 – February 2005	
4. TITLE AND SUBTITLE Army Enlisted Personnel Competency Assessment Program: Phase II Report			5a. CONTRACT OR GRANT NUMBER DASW01-03-D-0015/DO #5			
			5b. PROGRAM ELEMENT NUMBER 622785			
6. AUTHOR(S) Knapp, Deirdre J. and Campbell, Roy C. (Editors) (Human Resources Research Organization)			5c. PROJECT NUMBER A790			
			5d. TASK NUMBER 104			
			5e. WORK UNIT NUMBER			
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Human Resources Research Organization 66 Canal Center Plaza, Suite 400 Alexandria, VA 22314			8. PERFORMING ORGANIZATION REPORT NUMBER			
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) U. S. Army Research Institute for the Behavioral & Social Sciences 2511 Jefferson Davis Highway Arlington, VA 22202-3926			10. MONITOR ACRONYM ARI			
			11. MONITOR REPORT NUMBER Technical Report 1174			
12. DISTRIBUTION/AVAILABILITY STATEMENT Approved for public release; distribution is unlimited.						
13. SUPPLEMENTARY NOTES Contracting Officer's Representatives and Subject Matter POC: Peter Greenston and Tonia Heffner						
14. ABSTRACT (<i>Maximum 200 words</i>): <p>In the early 1990s, the Department of the Army abandoned its Skill Qualification Test (SQT) program due primarily to maintenance, development, and administration costs. This left a void in the Army's capabilities for assessing job performance qualification. To meet this need, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) instituted a 3-year program of feasibility research related to development of a Soldier assessment system that is both effective and affordable. The PerformM21 program has two mutually supporting tracks. The first focuses on the design of a testing program and identification of issues related to its implementation. The second track is a demonstration of concept – starting with a prototype core assessment targeted to all Soldiers eligible for promotion to Sergeant, followed by job-specific prototype assessments for several Military Occupational Specialties (MOS).</p> <p>The present report describes the second year of the PerformM21 program, in which a core examination was pilot tested and prototype test content was developed for five MOS. Further consideration was also given to program design features (e.g., delivery models, test frequency). Program design considerations include substantial attention to ways in which technology could be used to support the program and issues associated with the successful application of such tools.</p>						
15. SUBJECT TERMS Behavioral and social science Personnel Job performance measurement Manpower						
SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT Unlimited	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON Ellen Kinzer Technical Publications Specialist (703) 602-8047	
16. REPORT Unclassified	17. ABSTRACT Unclassified	18. THIS PAGE Unclassified				

Standard Form 298

Technical Report 1174

**Army Enlisted Personnel Competency Assessment
Program: Phase II Report**

Deirdre J. Knapp and Roy C. Campbell (Editors)
Human Resources Research Organization

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

U.S. Army Research Institute for the Behavioral and Social Sciences
2511 Jefferson Davis Highway, Arlington, Virginia 22202-3926

January 2006

Army Project Number
622785A790

**Personnel Performance and
Training Technology**

Approved for public release; distribution is unlimited.

Acknowledgements

U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) Contracting Officer Representatives (COR)

Dr. Peter Greenston and Dr. Tonia Heffner of ARI served as co-COR for this project, but their involvement and participation went far beyond the usual COR requirements. Their contributions and active input played a significant role in the production of the final product and they share credit for much of the outcome. Of particular note are their activities in conveying information about the project in briefings and presentations to Army Leadership on many important levels.

The Army Test Program Advisory Team (ATPAT)

The functions and contributions of the ATPAT, as a group, are documented in this report. But this does not fully reflect the individual efforts that were put forth by members of this group. Project staff is particularly indebted to Sergeant Major Michael Lamb from the U.S. Army Training and Doctrine Command (TRADOC) who serves as the ATPAT Chairperson.

The other current individual members of the ATPAT are:

SGM John Cross
CSM George D. DeSario
SGM (R) Julian Edmondson
CSM Dan Elder
CSM (R) Victor Gomez
SGM John Griffin
SGM John Heinrichs
SGM (R) James Herrell
SGM Enrique Hoyos
CSM Nick Piacentini

SGM David Litteral
SGM Michael Magee
SGM Tony McGee
SGM John Mayo
SGM Pamela Neal
CSM Doug Piltz
SGM (R) Gerald Purcell
CSM Robie Roberson
CSM Otis Smith Jr
MSG Matt Northen

ARMY ENLISTED PERSONNEL COMPETENCY ASSESSMENT PROGRAM: PHASE II REPORT

EXECUTIVE SUMMARY

Research Requirement:

The Army Training and Leader Development Panel NCO survey (Department of the Army, 2002) called for objective performance assessment and self-assessment of Soldier technical and leadership skills to meet emerging and divergent Future Force requirements. The Department of the Army's previous experiences with job skill assessments in the form of Skill Qualification Tests (SQT) and Skill Development Tests (SDT) were effective from a measurement aspect but were burdened with excessive manpower and financial resource requirements.

Procedure:

The U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is conducting a 3-year feasibility effort to identify viable approaches for the development of a useful yet affordable operational performance assessment system for Army enlisted personnel. Such a system would depend on technological advances in analysis, test development, and test administration that were unavailable in the previous SQT/SDT incarnations.

The ARI project (known as PerformM21) is being conducted with support from the Human Resources Research Organization (HumRRO) and entails three phases:

- Phase I: Identify User Requirements, Feasibility Issues, and Alternative Designs
- Phase II: Develop and Pilot Test Prototype Measures
- Phase III: Evaluate Performance Measures and Make System Recommendations

The objective of Phase I was to identify issues that the overall recommendation needs to take into account for a viable, Army-wide system (R. C. Campbell & Knapp, 2004). Phase I also produced a rapid prototype assessment covering Army-wide "core content" with associated test delivery and test preparation materials (R.C. Campbell, Keenan, Moriarty, Knapp, & Heffner, 2004).

In Phase II, the subject of the present report, the research team (a) pilot tested the core competency assessment, (b) developed competency assessment prototypes for five Military Occupational Specialties (MOS), and (c) explored issues further to develop more detailed recommendations related to the design and feasibility of a new Army enlisted personnel competency assessment program. In Phase III, the MOS tests will be pilot tested and a cost and benefit analysis of a notional Army program will be conducted.

Findings:

The prototype Army core assessment was successfully administered to over 600 E4 Soldiers. The test was web-based and delivered primarily through Army Digital Training Facilities. Although there were some technology-related delivery issues, the test itself, as well as the Test Preparation Guide, was well received by Soldiers. We also identified five MOS for which we collected job analysis information and developed prototype assessments. Assessment methods include job knowledge tests enhanced with advanced graphics features, situational judgment tests, and simulations. Finally, we extended our consideration of alternative program design features and technology issues that began in Phase I.

Utilization and Dissemination of Findings:

The core assessment work conducted in Phases I and II has resulted in lessons learned and a test item bank suitable for incorporation into an operational test program. The prototype MOS assessments will be pilot tested in Phase III of the PerformM21 research program. The program design and technology issues and recommendations are intended to help Army leaders make informed decisions about competency assessment. They will also be incorporated into a cost and benefit analysis to be conducted in Phase III.

CONTENTS

	Page
Chapter 1: PerformM21 Research Program Overview	1
Roy C. Campbell and Deirdre J. Knapp	
Introduction.....	1
NCO21 Research Program.....	2
PerformM21 Research Program	2
Supporting Groups	4
Research Approach: Integrating Process and Results.....	6
Guiding Principles	8
Overview of Report.....	9
Chapter 2: Army Assessment Program Design and Support Requirements	10
Deirdre J. Knapp and Roy C. Campbell	
An Overall Vision.....	10
Supporting Structure and Functions.....	14
Collect/Update Job Analysis Information.....	18
Prepare/Update Training and Doctrine	21
Design, Develop, and Update Assessments.....	22
Scheduling, Records Management, and Communication with Examinees	28
Deliver Assessments	29
Scoring and Score Reporting	30
Program Evaluation	32
Summary	33
Chapter 3: Pilot Test of the Army-Wide Examination.....	34
Karen O. Moriarty, Tonia Heffner, Roy C. Campbell, Huy Le, and Deirdre J. Knapp	
Background	34
Method	34
Soldier Reactions	39
Core Job Knowledge Exam Item Analyses	40
LeadEx Analysis	45
Subgroup Differences	45
Pilot Test Products	48
Recommendations for an Operational Test.....	48
Chapter 4: Technology Issues and Options	55
Roy C. Campbell, Jeffrey A. Barnes, and Shelly West	
Introduction.....	55
General Issues	55
Test Development and Delivery Software.....	58
PerformM21 Software: History and Experience.....	61
Test Delivery.....	63
Administrative Support.....	70
Summary and Conclusions	72

CONTENTS (continued)

	Page
Chapter 5: Development of Prototype MOS Assessments	73
Deirdre J. Knapp, Karen O. Moriarty, Roy C. Campbell, Chad Van Iddekinge, Lee Ann Wadsworth, Alicia Sawyer, Masayu Ramli, Andrea Sinclair, and Carrie Noble	
Introduction.....	73
Job Analysis and Test Design	76
Test Development	82
Operational Appraisals: Reviews of the Target MOS	107
Summary	121
Chapter 6: Summary and Next Steps	124
Deirdre J. Knapp and Roy C. Campbell	
PerformM21 Accomplishments to Date	124
PerformM21 Phase III.....	124
An Army-Wide Core Competence Assessment.....	125
Technical (MOS-Specific) Competence Assessment	125
Closing Comments.....	126
References.....	127

List of Appendices

Appendix A - DCAP Test Preparation Guide.....	A-1
Appendix B - Sample Soldier Feedback Report	B-1
Appendix C - Test Item Development Handbook	C-1
Appendix D - Job Analysis Ratings.....	D-1

List of Tables

Table 1. Major Design Features.....	10
Table 2. Phased Implementation Alternatives	14
Table 3. Assessment Methods.....	23
Table 4. Core Exam Pilot Test Locations	36
Table 5. Summary of Technological Issues	38
Table 6. Soldiers' Ratings of Their Performance	40
Table 7. Soldiers' Ratings of Item Effectiveness.....	40
Table 8. Estimated Weights for the Non-Traditional Items.....	42
Table 9. Distribution of Points and Items in the Core Item Set.....	43

CONTENTS (continued)

	Page
Table 10. Summary Item Statistics for Core Item Set based on Full Sample.....	43
Table 11. Descriptive Statistics of Core Item Set.....	44
Table 12. Correlations among Job Knowledge and LeadEx Scores.....	45
Table 13. Subgroup Performance Differences on the Core Job Knowledge Item Set.....	46
Table 14. Subgroup Performance Differences on the LeadEx Items.....	46
Table 15. MOS Category Differences on the Core Job Knowledge Item Set and LeadEx	47
Table 16. Correlations Between Demographic Variables and Test Performance	47
Table 17. Top-Down Rankings of First Aid Tasks.....	49
Table 18. Summary Item Statistics for Each Content Domain.....	50
Table 19. Current Bandwidth Capability Comparisons.....	67
Table 20. Population Estimates by Paygrade – Active and Reserve Components	68
Table 21. Select21 MOS	74
Table 22. PerformM21 Target MOS.....	75
Table 23. Summary of Site Visits and Subject Matter Expert Participation	76
Table 24. Assessment Methods by MOS.....	78
Table 25. Critical Incident Categories for 68W.....	87
Table 26. Features of Path and Open Simulations.....	97
Table 27. Design of the SACMS-VT.....	119

List of Figures

Figure 1. Outline of PerformM21 needs analysis organizing structure.....	7
Figure 2. Assessment program supporting structure and functions.....	16
Figure 3. Criterion- and norm-referenced tests, as defined by TRADOC Pamphlet 350-70-5 (p.18).24	
Figure 4. Major test development and maintenance activities.....	25
Figure 5. Presentation order.....	37
Figure 6. Screenshot of navigation tools for Soldiers taking the core examination.	52
Figure 7. Test administration requirements model.	56
Figure 8. Excerpt from the 68W STP Table of Contents.....	83
Figure 9. Excerpt from the TM 9-8000 Table of Contents.....	84
Figure 10. Content validation ratings questions.	85
Figure 11. SimMan® at Fort Sam Houston.....	95
Figure 12. Screenshot of 14E simulation.....	99

ARMY ENLISTED PERSONNEL COMPETENCY ASSESSMENT PROGRAM: PHASE II REPORT

CHAPTER 1: PERFORMM21 RESEARCH PROGRAM OVERVIEW

Roy C. Campbell and Deirdre J. Knapp

Introduction

The Department of the Army is changing to meet the needs of the 21st century. Soldiers at all levels must possess the interpersonal, technical, and organizational knowledge, skills, and other attributes (KSAs) to perform effectively in complex technical, information-rich environments, under multiple and changing mission requirements, and in semi-autonomous, widely dispersed teams. The Army needs an integrated Soldier assessment system to support these demands.

In June 2000, the Chief of Staff of the Army established the Army Training and Leader Development Panel (ATLDP) to chart the future needs and requirements of the Noncommissioned Officer (NCO) corps. After a 2-year study, which incorporated the input of 35,000 NCOs and leaders, a major conclusion and recommendation was: "Develop and sustain a competency assessment program for evaluating Soldiers' technical and tactical proficiency in the military occupational specialty (MOS) and leadership skills for their rank" (Department of the Army, 2002).

In the early 1990s, the Army abandoned its Skill Qualification Test (SQT) program due primarily to maintenance, development, and administration costs. Cancellation of the SQT program left a void in the Army's capabilities for assessing job performance qualification. Re-instituting a new performance assessment system must address the factors that forced abandonment of the SQT. Since then, technological advances have occurred that can reduce the developmental and administrative burdens encountered with SQT and will play a critical role in a new performance assessment system.

To meet the Army's need for job-based performance measures, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) instituted a 3-year program of feasibility research, called Performance Measures for the 21st Century (PerformM21), to identify viable approaches for development of a Soldier assessment system that is both effective and affordable. This research is being conducted with contract support from the Human Resources Research Organization (HumRRO). The impetus to include individual Soldier assessment research in ARI's programmed requirements began prior to 2000 and was based on a number of considerations regarding requirements in Soldier selection, classification, and qualifications. For example, lack of operational criterion measures has limited improvements in selection and classification systems. Meanwhile, there were several significant events within the Army that reinforced the need for efforts in this area. The aforementioned ATLDP recommendation resulted in the Office of the Sergeant Major of the Army (SMA) and the U.S. Army Training and Doctrine Command (TRADOC) initiating a series of reviews and consensus meetings with the

purpose of instituting a Soldier competency assessment test. Ongoing efforts within the Army G1 to revise the semi-centralized promotion system (which promotes Soldiers to the grades of E5 and E6) also were investigating the use of performance (test)-based measures to supplement the administrative criteria used to determine promotion. Ultimately, the three interests (ARI, SMA/TRADOC, G1) coalesced and the ARI project sought to incorporate the program goals and operational concerns of all of the Army stakeholders, while still operating within its research-mandated orientation.

NCO21 Research Program

Prior to the start of the PerformM21 research program, another G1-sponsored, ARI-HumRRO effort specifically addressed changes to the semi-centralized promotion system. Called NCO21, this research focused on the identification of key KSAs for NCO performance (particularly as they apply to future Army requirements) and resulted in the subsequent development of new tools that could be incorporated into the promotion process (R.C. Campbell, Knapp, & Heffner, 2002; R.C. Campbell, J.P. Campbell, Knapp, & Walker, 2000; Knapp, Heffner, & R.C. Campbell, 2003; Knapp, McCloy, & Heffner, 2004; Knapp et al., 2002). These measures included temperament indicators, situational judgment exercises, a semi-structured interview process, and reweighting of existing semi-centralized administrative points allocations as well as applying differing metrics by pay grade and MOS. Preliminary validation, using a concurrent validation sample, indicated significant predictive achievements to the promotion system using the new measures. The measures that showed the most potential for implementation comprise the "Leadership Assessment Tool" (LAT), which is now being evaluated in a longitudinal validation. This evaluation involved administration of the LAT to approximately 1,000 E4 and E5 Soldiers in 2004. Post-promotion performance information, which will be used as criteria for further evaluation of the LAT, will be collected from the Soldiers' supervisors in 2005.

The NCO21 research program confined itself to promotion tools that could be applied Army-wide and which did not tap detailed task knowledge and skill. The PerformM21 research program is thus best viewed as a companion effort that addresses this gap. In other words, future improvements to the NCO promotion system would make use of the products and findings generated by both of these related lines of research.

PerformM21 Research Program

PerformM21 has three phases:

- Phase I: Identify User Requirements, Feasibility Issues, and Alternative Designs
- Phase II: Develop and Pilot Test Prototype Measures
- Phase III: Evaluate Performance Measures, Conduct a Cost-Benefit Analysis, and Make System Recommendations

Phase I of PerformM21 resulted in program design recommendations that included such considerations as how an Army assessment would be delivered, how assessments would be designed, developed, and maintained, and what type of feedback would be given. We also

developed a demonstration Army-wide core assessment test to serve as a prototype for the envisioned new Army testing program. This core assessment is a computer-based, objective test that covers core knowledge areas applicable to Soldiers in all MOS (training, leadership, common tasks, history/values). Phase I was completed in January 2004 and is documented in two ARI publications (R. C. Campbell, Keenan, Moriarty, Knapp, & Heffner, 2004; Knapp & R. C. Campbell, 2004).

Phase II of the PerformM21 program (which corresponds roughly to year two of the 3-year overall effort) had three primary goals:

- Conduct an operational pilot test of the Army-wide core assessment with approximately 600 Soldiers.
- Investigate job-specific competency assessments in up to five MOS.
- Continue to refine and to develop discussion and recommendations related to the design and feasibility issues established in Phase I.

In addition to the core elements of PerformM21 broadly outlined in the three Phases, there have been two related studies generated by requirements uncovered during the PerformM21 research. These are:

- The Self-Assessment Program: A study to determine the kinds of information Soldiers need to determine their overall readiness for promotion, including identification of strengths and weaknesses prior to testing (Keenan & R.C. Campbell, 2005).
- Deployment Skills Identification: A study to determine new or refocused skills and tasks associated with operations in Iraq and Afghanistan and to include those requirements in an Army core assessment program.

These additional studies (completed in January 2005 and to be completed in December 2005, respectively) are reported on separately.

The PerformM21 research program is best viewed as having two mutually supporting tracks. The first track is essentially conceptualization and capture of issues, features, and capabilities in Army testing design recommendations. The second track is to develop and administer prototype operational tests and surrounding supporting and preparation materials. These prototypes include both the Army-wide core assessment and some selected MOS tests. These are intended to reflect, inasmuch as possible, design recommendations for the future operational assessment program. Experiences with the prototypes will, in turn, influence elaboration or modification of the operational program design recommendations as they develop during the course of the 3-year research program.

Much of the research involves refinement of issues over the life of the project so, in that regard, the Phases are continuous and the Phase products and reports are building blocks. Also the objective is to close the gap between research and potential operational implementation. As that gap narrows, factors such as Army policies, organizational structure, resources, and real-life

restrictions become more critical. Final considerations will include a cost-benefits analysis, to be conducted in 2005 as part of Phase III.

Supporting Groups

The Army Test Program Advisory Team (ATPAT)

Early in Phase I, we constituted a group to advise on the operational implications of Army assessment testing, primarily as part of the needs analysis aspect of the project. Simultaneously, this group took on a role as Test Council for the Army-wide core assessment. Subsequently, the group has become an all-around resource for all matters related to potential Army testing. This group is called the Army Test Program Advisory Team (ATPAT) and it has the following characteristics:

- It is made up of NCOs, mostly in the Master Sergeant (E8) and Sergeant Major (E9) levels.
- It includes representatives from TRADOC, HQ, Forces Command (FORSCOM), Combined Arms Center (CAC), Center for Army Leadership (CAL), Army Training Support Center (ATSC), Army G1, Sergeant Major Academy (USASMA), and the Office, Sergeant Major of the Army. During Phase II, we added specific organizational representation from TRADOC schools corresponding with the five MOS that were selected for exploration and development (Air Defense, Armor, Medical, Military Police, Ordnance).
- It includes representatives from the Reserve force including from HQ, Army National Guard Bureau (ARNGB), HQ, Army Reserve Command (USARC), and unit representatives from the 95th Division (Institutional Training), 653rd Area Support Group (ASG), and the 78th Division (Training Support).
- It is co-chaired by a Sergeant Major endorsed by the ATPAT. The chair is responsible for approving agenda items for meetings and for serving as point of contact for the ARI/HumRRO project team.
- It has a flexible membership. Although there is a solid core ATPAT group, there have thus far been over 40 individual representatives to the ATPAT. The ATPAT usually meets quarterly, however individuals and smaller groups from the ATPAT work with project staff on a regular basis addressing specific issues and problems.

The ATPAT serves several purposes. First, it provides the primary input for the continuing refinement of the needs analysis requirements of the project, primarily by providing insight into operational implications and real-world feasibility of the program. Second, it serves as the oversight group for development of the Army-wide core test and of the MOS tests as well as a resource in identifying and developing content for the tests. Additionally, the ATPAT is a working group that provides product reviews, subject matter expertise, and, as needed, assistance in the process of developing prototype instruments and trial procedures. An additional benefit of the ATPAT is to serve as a conduit to explain and promote the PerformM21 project to various Army agencies and constituencies.

Specifically, we have looked to the ATPAT to provide guidance in four areas:

- *Utilization Strategies* – How the test will be used, defining and limiting the scope of the program. That is, whether and how it will be used in personnel management, promotion, career development, training, readiness, retention, and transition.
- *Implementation Strategies* – Identifying the steps to implementing, maintaining, and growing an Army test, short- and long-term goals, and organizational implications to be considered in phased implementation.
- *Operational Strategies* – Identifying the considerations that must be taken into account to operationalize an Army testing program (for developers, administrators, and users).
- *External Considerations* – How an Army test program will fit in with other agendas such as self-development, unit training, the NCO Education System (NCOES), deployments, the NCO Evaluation Report (NCOER), Table of Distribution Allowances (TDA) staffing, transition, Soldier tracking and assignment, Future Force, and training publications and updates.

The ATPAT has been extremely helpful in discussions in all of these areas. Since the project started in 2002, there have been significant changes in the operational posture of the Army, primarily in support of the expeditionary force and personnel turbulence associated with an Army at war. The ATPAT has played a significant role in ensuring the project maintains relevance under these changing conditions.

Joint Services Working Group

Both the U.S. Air Force and U.S. Navy have large-scale promotion test programs. Early in Phase I, project staff visited with both of these programs to exchange information. While neither program provides an exact model that matches the Army's needs, both programs have extensive experiences that apply to the current research.

Subsequent to these early visits, we have maintained professional contacts within the other services to continue the exchange of testing ideas. In 2004, these contacts evolved into the establishment of an informal and unofficial Joint Services Working Group to continue this relationship. The group, which has been expanded to include representatives from the U.S. Coast Guard and the U.S. Marine Corps, has established a Joint website and conducts teleconferences on an as-needed basis. PerformM21 experiences with web-based computer testing has been a topic of particular interest, given it is an area that the Navy and Air Force are interested in pursuing as well.

Core Technology Group

From the outset of the project, it was our *a priori* assumption that advances in technology would be key to designing an assessment program that will be feasible and affordable. From the start, all test development and administration has been technology centered. As we gained more actual experience during the pilot testing of Phase II, we discovered that technology issues were both significant and continuing. Much progress has been made and we are committed in our

resolve that technology will eventually result in an efficient, cost effective program. But a technology approach to large scale distributed testing involves many facets and there is considerable discovery learning that is occurring during the pilots.

Early on, we organized a team of HumRRO personnel (and augmented by outside resources as necessary) with particular knowledge and experience of technology as it applies to large-scale assessment to learn about the Army's existing capabilities and craft recommendations related to this area. We call this team the Core Technology Group and it has grown in size and in focus as the project has evolved. Initial activities of this group were documented in Appendix E to the Phase I Army-wide prototype test report (R.C. Campbell et al., 2004) and the technology issues (test development and delivery software, delivery portal, web-delivery support requirements) are updated in Chapter 4 of this report. An important role of this group is to track testing requirements with existing and planned Army initiatives in the technology field. The Core Technology Group will continue to function as a resource throughout the project.

Research Approach: Integrating Process and Results

The PerformM21 project is complex. It encompasses separate but intertwined tracks for the Army-wide and MOS tests, three Phases over approximately 40 calendar months, many diverse sources and resources, both within and outside of Army channels, and an ever-changing Army operational posture. Outcomes include both hard products, such as prototype tests and data from Soldier tryouts, and analytic requirements such as lessons learned and implementation strategies. A key to keeping everything organized has been the *Needs Analysis Organizing Structure*. Developing this requirements-based structure was a major focus of Phase I and its importance has been reinforced in subsequent phases.

To structure the needs analysis process, project staff drafted a list of requirements for supporting an assessment program. Figure 1 lists the key components, and this organizing structure is more fully explained in the Phase I needs analysis report (Knapp & R.C. Campbell, 2004). This structure helped organize our thinking and suggested the questions we posed to those providing input into the process. We obtained input from several sources as we considered the issues, ideas, and constraints associated with each requirement listed in Figure 1. These included the following:

- The Army Testing Program Advisory Panel (ATPAT)
- Historical information about the SQT program and associated lessons learned
- Enlisted personnel promotion testing programs operated by the Air Force and the Navy
- Civilian assessment programs (e.g., professional certification and licensure programs)
- A review of automation and technology tools and systems
- Work completed under a related project – a Phase I Small Business Innovation Research (SBIR) grant¹

¹ In 2003, Job Performance Systems, Inc. (JPS) and HumRRO performed work under an SBIR Phase I program that was designed to complement the PerformM21 research program (Rosenthal, Sager, & Knapp, 2005). The primary product was a proposed methodology to use to produce realistic and cost-effective job performance assessments. This methodology was incorporated into the job analysis work performed to support the MOS test development as part of PerformM21 Phase II and as described later in this report.

-
- Purpose/goals of the testing program
 - Test content
 - Test design
 - Test development
 - Test administration
 - Interfacing with candidates
 - Associated policies
 - Links to Army systems
 - Self-assessment
-

Figure 1. Outline of PerformM21 needs analysis organizing structure.

This needs analysis organizing structure has been crucial to the project. It is the instruction manual for a prototype Army test system. The needs analysis process is dynamic; the requirements are revisited on a regular basis for updating. Although the main topics of the organizing structure remain unchanged, as more is learned through research, trials, and inputs from Army and systems experts, the content of the structure becomes more robust.

While the organizing structure provides the framework, the approach to the PerformM21 work has involved pulling together many diverse parts and products. Information is being obtained from many sources and in many different formats. New ideas are generated and old ideas are refined and reinforced. Some concepts do not survive tryout and some new procedures emerge. As we gather more information and gain more experience in producing tests and conducting test tryouts, we expand or revise our original thinking, as required. Updated information regarding related programs, such as the Army's Learning Management System (LMS) or changes to the NCOES, are factored into the framework. We are constantly integrating the many parts of the project, some occurring simultaneously but being worked separately, and some integrating old information with new efforts and outcomes.

A prime example of how the approach works is illustrated in the development of the prototype MOS tests as described in Chapter 5 of this report. We deliberately set out to try different methods and different tactics to see what would work and what would not, knowing that the process would appear somewhat chaotic while it was ongoing but that the inconsistent approach could uncover novel procedures and results. Looking only at the approach to any single MOS looks like just bits and pieces, but when all experiences are put together *post facto*, they make a more coherent whole. Such is the case with the entire project. Periodically, project staff pull together all the assembled pieces and results, assess the implications, and reaffirm, refine, or reformulate ideas and positions.

The process is cumulative, iterative, continuous, and collaborative. In the end, our goal is to provide the Army as much information and guidance as possible to facilitate the leadership's decision-making about a new Army test program. To do this, we need to marshal data, experiences, reasoned judgments, and a timely appreciation of influences and priorities within the Army and present them in logical, acceptable formats.

Guiding Principles

Early in the work, we collectively brainstormed what an Army test program should look like, at a very general level. The result was a set of guiding principles, which, like other parts of the project, have gone through some revisions and refinements. However, they still maintain their relevance midway through the project and should be kept in mind as guideposts and reminders throughout:

- *The Art of Assessment:* We know a great deal about how to develop good assessments – ones that are valid and psychometrically reliable and that make optimal use of advanced technology. What presents a particular challenge for the present situation is designing an assessment program that will be high quality, cost-effective, and practically supportable by the Army. If resources were not an issue, this would be relatively easy.
- *The Cost of Assessment:* As becomes more evident as discussions and planning continue, an effective assessment program that includes core and MOS-specific assessments for Soldiers at multiple pay grades will be an expensive undertaking. Generally speaking, however, start-up costs will be considerably higher than maintenance costs. Rolling out the assessment program in phases will make the upfront costs more manageable.
- *The Timing of the Assessment Program:* There are other significant advantages to growing and developing the assessment program over time. First, the program must develop the trust of Soldiers. While it is true that some examinees will always criticize the tests they have to take, the Army can accomplish a great deal by making sure that Soldiers learn to respect this new assessment program. Early problems can damage the program's reputation in ways that are virtually impossible to overcome once that damage is done. Our advice is to start slow and strong and manage expectations as much as possible. This means being clear that the program will grow and improve with experience and input from all stakeholders.
- *Buy-In to Assessment:* It is not just individual Soldiers who need to be convinced the assessment program is worth doing and doing well. The more that all supporting organizations and individuals are vested in and respect the idea, the more effective it will be. It will therefore be important to “market” the program to all stakeholders by informing them of program goals and plans while also educating them about the various considerations and constraints that go into developing a successful program. People invariably underestimate what it takes to develop and maintain a high quality assessment program and they often question the motives of those developing those programs.
- *Integration into the NCO Development and Promotion System:* The assessment work conducted under PerformM21 is only a part of a much larger picture. NCO development for the 21st century is a many-faceted program, encompassing changes in NCO education, self-development, utilization and assignment, and selection. The work done under the NCO21 project lays the groundwork for an overall transformation of the semi-centralized promotion system, of which the PerformM21 assessments would be a part. These related but independent efforts, within ARI and elsewhere, need to be melded into a coherent system to serve the Army and the NCO corps in the coming years.

- *Quality of the Assessment Program:* Finally, if the Army is serious about instituting an assessment program, it must be willing to support it without cutting corners. For as important as it is to have, a low quality assessment program would be worse than none at all. A poorly maintained program would be unfair to Soldiers and would not help the Army improve readiness. Commitment to quality is critical to the program's success.

Overview of Report

The remainder of this report discusses the status of the project and the results achieved as they stand at the conclusion of Phase II. Specifically, Chapter 2 provides a detailed update of the design features of the Army assessment program. Chapter 3 describes the Army-wide prototype pilot test administration, its results, and recommendations for operational testing. Chapter 4 provides updates on the critical technology issues and options, starting with a recap of the decisions made in Phase I and including lessons learned from the Army-wide test pilot. Chapter 5 is a comprehensive overview of the development of the prototype MOS assessments: job analysis, test design, and test development. Finally, Chapter 6 summarizes what we see as the major accomplishments to date, outlines plans for Phase III, and discusses briefly some of the other steps required to implement the envisioned assessment program.

As we move through Phase III of the PerformM21 research program, the vision of the new program will no doubt evolve and become more detailed. In the meantime, this report is intended to be a snapshot that captures where the concept stands at this moment in time.

CHAPTER 2: ARMY ASSESSMENT PROGRAM DESIGN AND SUPPORT REQUIREMENTS

Deirdre J. Knapp and Roy C. Campbell

The purpose of this chapter is to provide an overall vision of a new Army competency assessment program, discuss the functions the Army would need to incorporate into its structure and systems to support such a program, and review some of the design issues associated with the major elements of the program. Many of these issues will be discussed further in subsequent chapters that describe our experiences thus far developing and delivering prototype assessments.

An Overall Vision

Major Design Features

While nothing about the competency assessment program will be certain until the Army determines if and how the program will be implemented, Table 1 lists some of the recommended basic features of the assessment program. This vision is based on the input of the ATPAT, guidance from the Office of the SMA, and the knowledge and experience of the PerformM21 project staff.

Table 1. Major Design Features

- All Soldiers in pay grades E4 through E7 will be included in the program
 - Scores will be used to support promotion decisions
 - The assessment program will be the same for all components of the Army
 - There will be an Army-wide core competency test and/or MOS-specific tests
 - Assessments will be computer delivered in a proctored environment
 - Each test will be administered during a test window period each year
 - Soldiers will be given adequate tools to prepare for the tests
 - Scores will be valid for a period (e.g., 2-3 years)
-

Target Pay Grades

Including virtually all Soldiers being considered for promotion into or within the NCO ranks (pay grades E4-E7) was an early and fixed recommendation from the SMA. This is also consistent with the recommendations from the ATLDP (2002) and reflects the view that the NCO corps will be strengthened by ensuring that all Soldiers in NCO leadership positions are periodically required to demonstrate competence on critical job requirements in a way that can be evaluated objectively.

Use Scores to Support Promotion Decisions

Although the assessment program will encourage self-development, using the assessment for Soldier feedback only (as opposed to using it for making personnel decisions) is not recommended. Maximum motivation for Soldiers to learn what they need to know will only

come if test scores make a difference in their career status. Thus, a critical recommended feature of the new assessment program is that scores be incorporated into the NCO promotion decision process. It is important to bear in mind that the primary purpose of the assessment program is to help ensure Soldiers know how to perform their jobs, not for them to do well on a test, per se.

The Army uses basically two different NCO promotion programs: The semi-centralized promotion system to ranks of Sergeant (E5) and Staff Sergeant (E6) and the centralized promotion system which promotes to Sergeant First Class (E7), Master Sergeant (E8), and Sergeant Major (E9). The semi-centralized system is a points based program, which (currently) uses administrative points supplemented by commander's points; and a local board points award system, to determine promotion standing within the individual Soldier's MOS. The centralized system involves selection by centralized boards based on record reviews and following specific instructions prepared for each selection board.

The semi-centralized system is the easiest to revise and is the obvious start point for the introduction of a test component as a part of the promotion system. It is also a system that is primed for review and revision, which, except for some superficial adjustments of point emphasis, is basically the same system that existed in the 1970s. The centralized promotion system is a more complex issue. Because it is less structured, the introduction of promotion test criteria needs more review and consideration as to how it should fit within either the existing or a revised centralized promotion system. One fairly simple adoption strategy would be to just provide test results as another element for the selection board to consider.

The Same Program for All Components of the Army

To maximize the program's potential to improve force readiness, it is important that the program apply to all components of the force equally. This means that the program's design must work effectively for both the Active Component and the Reserve Components (the U.S. Army Reserve [USAR] and the Army National Guard [ARNG]) and the same performance expectations should apply across components. For example, all pay grade E5s in the 63B MOS should take the same test, regardless of their component or duty status. A related consideration emphasized by the ATPAT is that the assessment program (and the NCO promotion system as a whole) must be sustainable during periods of heightened deployment activity.

At the same time, differences between the Active and Reserve Components must be recognized, primarily in areas such as accessibility and availability. These differences will mostly affect administration requirements, (e.g., test windows) and in eligibility criteria and scheduling as well as reporting and analysis. But the expectation is that test content and presentation would be the same across components.

Assess Core and/or MOS Competence

There is not yet a final consensus on whether the new assessment program should include a core competency test suitable for all Soldiers regardless of their MOS and/or MOS-specific assessments. Ideally, both core and MOS assessments would be included to ensure coverage of all relevant job requirements. It is also true, however, that more tests result in higher testing

costs. Guidance from SMA (R) Tilly in 2003 was to begin with a core assessment and possibly add MOS assessments later. He suggested the core assessment should include common tasks, leadership, training, history, and Army values and these are the areas covered by the prototype examination we developed in Phase I (R.C. Campbell et al., 2004). Certainly, a test program that covered only core competencies would be least expensive to support. It would mean having at most four exams (i.e., one exam for each of four pay grades – E4, E5, E6, E7), and likely fewer than that (it is likely that the test content differences will decrease as the higher pay grades are added). However, the picture quickly expands when you consider adding upwards of 180 MOS-specific assessments, with multiple versions of each to reflect different skill levels.

Some might argue that MOS assessment is more critical to force readiness (given the relative emphasis in training of core versus MOS-specific content), but the much greater cost associated with MOS-specific testing is undeniable. One reasonable possibility is to gradually phase in MOS-specific testing and to not require all MOS to participate. This might initially strike some as unfair, but the fact is that promotion decisions are made within MOS. It is already more or less difficult to get promoted in different MOS depending on a variety of factors, most notably the supply versus demand for NCOs at successive ranks. Indeed, for MOS that need to promote as many Soldiers as possible to fill manning requirements, additional promotion testing will have relatively little effect on readiness because most Soldiers who meet the most minimal criteria will be promoted regardless of their test performance.

Computer Based Testing

Computer based testing is another key feature of the program and one of many that distinguishes it from the old SQT program. Computer based testing has several benefits. In particular, it (a) avoids many of the security hazards and costs associated with a large-scale paper-based test program (e.g., printing and keeping track of thousands of test booklets), (b) allows for the use of state-of-the art assessment formats, and (c) likely has more credibility with Soldiers than paper-based tests. To maintain test security, the computerized tests will need to be delivered in secure facilities under the supervision of test proctors.

Test Administration Windows

For test security reasons, it would also be ideal to administer the same assessment at the same time to all examinees. This is not feasible, however, particularly when one considers the needs of deployed Soldiers and those Soldiers in the Reserve Components. Thus, the program should allow Soldiers to take the test within a specified window of time (e.g., 3-4 months). The shorter the test window, the less convenient it will be for the Soldiers. Longer test windows will cost more to support because of the greater exposure of the test questions. Note that the test window does not need to be the same for every test. For example, the core examination for E4 Soldiers could be administered in January through March and the core examinations for E5 through E7 Soldiers could be administered in August through October. Staggering test windows would have the advantage of spreading out the work required to support testing relatively evenly rather than concentrating it during a single time of the year.

Soldier Test Preparation

Soldiers will need to be supported in their efforts to prepare for the assessment, so this is included as a basic feature of the envisioned program. This support includes tools such as a test preparation guide, self-assessment exercises, and up-to-date training manuals. In addition to providing tools, serious consideration will need to be given to policies associated with the time needed for test preparation. A key question is the extent to which test preparation must be done on a Soldier's own time versus during duty time. In any case, the ATPAT has been firm in its belief that the assessment program should motivate Soldiers to "go back to their books."

Require Testing Every 2-3 Years

Finally, not requiring Soldiers to take a test every year would have important advantages over having an annual testing requirement. It would save money by significantly reducing the number of tests delivered each year. Given the current size of the Army, annual testing for all E4 through E7 pay grade Soldiers would mean delivering on the order of 500,000 tests per year. Reducing that figure by a quarter or a half would save considerable resources. Testing each Soldier every 2-3 years instead of every year also reduces the test preparation burden on Soldiers while still ensuring that they periodically demonstrate competence through an objective assessment.

Phased Implementation

We recommend that the Army's new assessment program be implemented in phases because this will (a) spread out start-up costs, (b) minimize the likelihood of start-up problems, and (c) allow lessons learned in the early phases to influence improvements to the program as it expands. It is very important that stakeholders have faith in the program, so minimizing the possibility of system failures at the beginning of the program will go a long way towards ensuring its credibility and acceptance. Cost considerations are also paramount, as the program will be expensive to implement and maintain. Note that an analysis of expected costs will be carried out as part of Phase III of this research program.

The actual plan for phasing in the assessment program will be determined by Army leadership priorities. Table 2 summarizes some possible strategies. Specifically, the phases could be based on assessment type, Soldier pay grade, Army component, or some combination of these. For example, the program could start with an Army-wide core competency assessment suitable for enlisted Soldiers at the Specialist/Corporal level (E4). The program could then expand to include core assessments for higher enlisted grade levels (E5 through E7) and to include MOS-specific assessments.

Wherever the program starts, we recommend that the first operational (as opposed to research) administration of the examination be a dry run, with scores not used for promotion points. This will help calm those Soldiers who are anxious about the idea of testing and allow the Army the opportunity to work out unexpected difficulties in large-scale test delivery before the test scores "count."

Table 2. Phased Implementation Alternatives

Based on type of assessment

- Start with core competency assessment
- Add MOS assessments as they become available

Based on pay grade

- Start with assessments at one pay grade (e.g., E4)
- Add additional pay grades

Based on component

- Start with one component (e.g., the USAR)
 - Add the other components later
-

Supporting Structure and Functions

Oversight and Coordination

There are two critical parts to defining the supporting structure for an Army test system. The first is to define the functions that must be performed and the second is to identify the entity that will have organizational responsibility and staffing to support these functions. The latter is largely a matter of Army policy and will be made in due course by Army leadership concurrent with decisions about operational testing. But organizational structure and responsibility is a key ingredient to a successful program and must be planned.

Historically, until about 1974, promotion testing was a function of the Army personnel system (Army G1). With the introduction of the SQT, responsibility for development and administration of testing was transferred to Army operations and training (G3). Considerable discussions have taken place within the ATPAT concerning the Army level of responsibility for a future test system. Given the relationship of testing to job requirements, performance documentation, training support, and operational implications, the consensus of the ATPAT is that the test program be a functional responsibility of the Army Deputy Chief of Staff G3 with TRADOC as its implementing agency. This position was reached with the full appreciation that the primary purpose of such a test system would be to input results into the NCO promotion system. The staffing responsibility for testing within TRADOC is consistent with concurrent initiatives regarding changes to NCOES and the development of the NCO Corps Vision Statement.²

Staffing needs to be established at a sufficiently high organizational level to be able to establish testing policy, initiate directives, and provide oversight and coordination. An Army assessment program will be most effective if it resides in an organization whose sole function is assessment. Quite likely, this will require a new organization—chartered, staffed, and resourced to carry out the Army testing mandate. The new organization would have responsibility for

² Based on the Draft Concept Paper: *Growing the 21st Century NCO Corps Transforming to a Relevant and Innovative NCO Development Strategy* (2004). This Concept Paper is still in development and not an official TRADOC paper but has been made available to the PerformM21 project for special consideration.

Army testing doctrine and policy and implementation of the testing program throughout the Army as well as overall test administration and utilization policy. It would be directly responsible for the development and maintenance of the Army-wide competency assessment test. Organizational authority and responsibility would encompass all components (Active and Reserve) and would include all MOS proponents as well as affecting directly most major commands and Army forces. Especially initially, a revived testing program will require some cultural adjustment within the Army; an appropriate level of organizational authority and certainty can have an important impact on the effectiveness and acceptance of the program. This Army organization should be at the Directorate level, headed by a Colonel (O6) and sufficiently staffed with both test psychologists and military representation.

For an Army test program to move forward, it is paramount that the Army establish ownership and assign organizational responsibility. But there are competing issues that must be recognized as well, including current priorities and demands on an Expeditionary Army. Timing on when to involve Army leadership in information and decision briefings is a sensitive concern that requires insightful analysis of current factors and viewpoints. We continue to rely on the input from the ATPAT for guidance on when and where to enter the testing issues into the leadership agenda. Meanwhile it is crucial that we continue to advance knowledge and planning on the myriad of issues surrounding Army testing in anticipation of a critical appraisal of initiation of an operational program.

Figure 2 illustrates both the functions and a notional organization required to support an assessment program. Unlike the former SQT program, the new program is likely to have a core assessment that does not have an existing single proponent. Moreover, it will be necessary for a cost-effective program to have a high degree of coordination and even some measure of control among MOS proponents. Such a group would meet periodically to determine strategies for sharing resources and lessons learned. The group would also help ensure that differences in MOS assessment policies and procedures make sense in light of MOS differences rather than being haphazard in nature across the Army. Finally, this group would be a vehicle for coordinating with the other military services and the civilian testing community to identify and explore avenues of mutual support.

As depicted in Figure 2, the "Army Assessment Program Director" would be the officer responsible for overall Army test program policy in development and administration. This office would also be responsible for overseeing and coordinating policy within the various MOS proponents and for coordinating test use and implementation among the many Army agencies and commands.

A Council of Sergeants Major would be responsible for making recommendations related to the design, content, and policies associated with the core assessment program. In the development of the prototype Army core examination, this function has been served by the ATPAT. It is reasonable to think the ATPAT could evolve into the standing Council of SGMs with suitable guidelines for membership, composition, and responsibilities. If the Army elected to institute either core competency assessment or MOS-specific assessment, but not both, oversight and coordination requirements would be simplified.

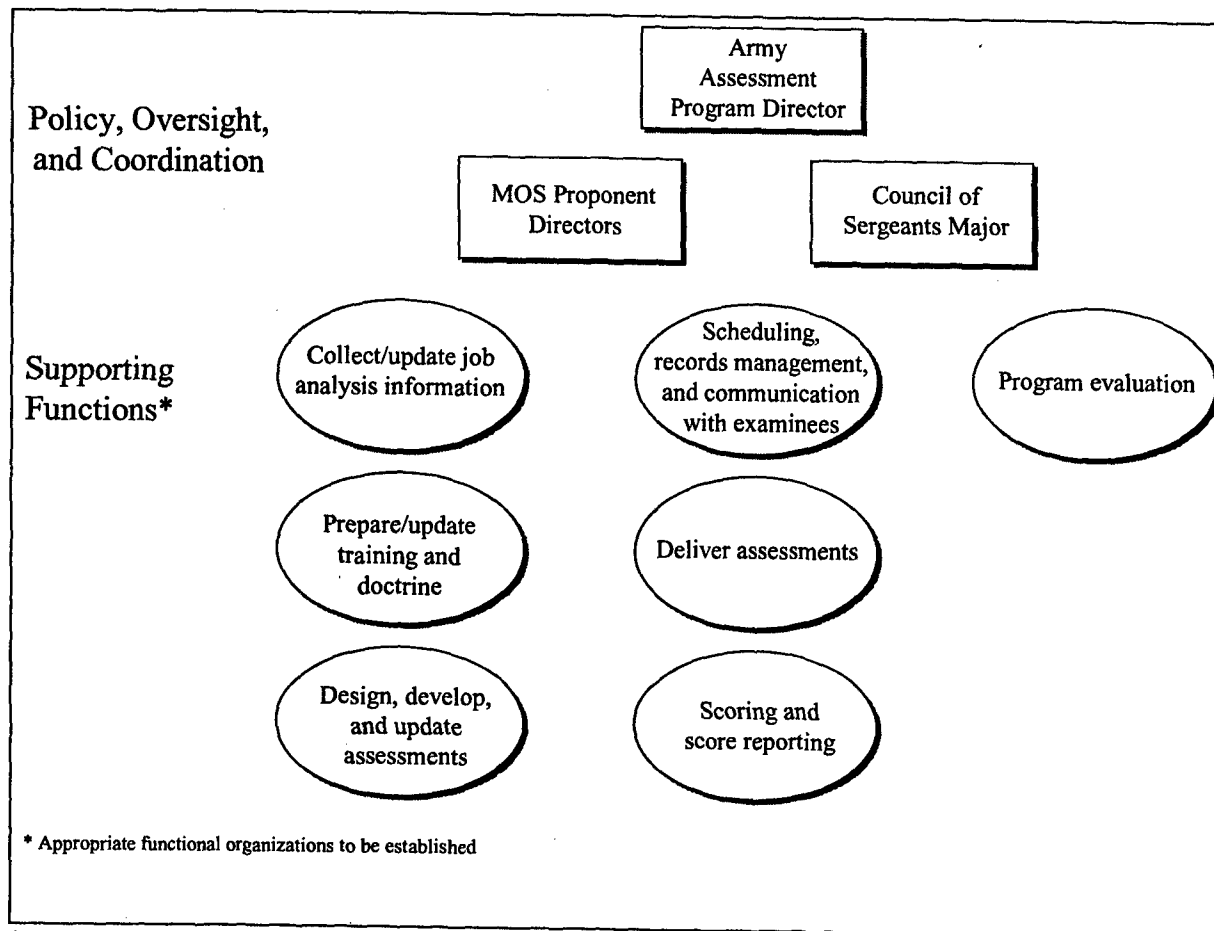


Figure 2. Assessment program supporting structure and functions.

Supporting Functions

The lower part of Figure 2 notes the major functions that are required to support an effective assessment program. They are derived from what is needed to (a) develop and maintain the assessments (i.e., job analysis data; training and job aids on which to base test content; people and systems to design, develop, and update the assessments), (b) administer the assessment program (e.g., scheduling and delivering the assessments, scoring and score reporting), and (c) evaluate the program.

Policies

The translation of test results to promotion decisions involves issues of scoring, management, weighting, standardization, and equating. There is a great deal of flexibility possible in the application of test results as well as many considerations and cautions in how testing results are applied. It is crucial that personnel policy decisions be based on sound incorporation of testing science and experience and that there be a mutual understanding of both the requirements of the promotion program and the products and application of the test program.

A final area with personnel policy implications is in the area of test security. Both the Navy and Air Force have clear-cut personnel policies and command emphasis on test compromise, test control, and cheating. Both services investigate violations aggressively and have successfully prosecuted individuals under the Uniform Code of Military Justice (UCMJ) for willful or deliberate breaches. The Army needs to identify, implement, and publicize a similar approach.

Staffing

Under the SQT program, tests were designed, developed, and delivered primarily by Active duty personnel. The significant demands of supporting a testing program while effectively handling other operations, however, was a significant factor in the demise of this program (see Appendix B in Knapp & R.C. Campbell, 2004). Moreover, today's Army is much smaller than it was back then and has an ever-increasing number and diversity of missions (Ford et al., 2001; Sager et al., 2004). These factors make it even more important to ensure Soldier readiness while making it less feasible for Active duty personnel to handle the brunt of the supporting manpower requirement.

As we will note in subsequent sections, it would be preferable to have some activities needed to support the testing program performed by testing professionals or at least with their close oversight. The typical model used by the Army and the other U.S. military services is to prescribe exactly how things should be done in a procedural manual (e.g., how to conduct a job analysis or develop certain types of tests), thus allowing anyone to do the required activities. There are at least two flaws in this model. The first is that test design and development can be highly proceduralized, but that does not mean that the results will be high quality. Involvement by testing professionals is needed to help ensure the products (e.g., test questions) that result from following those prescribed procedures meet quality standards. The second issue relates to modification of procedures. The science underlying the testing profession is not static. Using the input of testing professionals, the Army could establish a high quality testing program, replete with detailed procedures for designing, developing, scoring, and maintaining tests. But those procedures will become outdated and could easily become eroded as practical considerations lead to shortcuts that compromise the original intent of the procedures. Thus the program and its procedures should be routinely scrutinized and modified to reflect advances in the science of testing or problems identified in the program. To the extent the program is run entirely by "implementers" instead of those who have the academic and experiential background to identify the need for improvements, the long-term quality of the testing program is likely to suffer.

To address these concerns, we suggest that some functions (e.g., test development) largely be performed with contractor support. Contractors vying for this work would likely hire cadres of retired NCOs whom they could teach to collect job information and develop tests. The contractors would also be expected to employ testing professionals with the expertise to oversee this work and to handle the more complex aspects of the work (e.g., determining how to score complex simulation tests). Of course, the Army should also have sufficient in-house expertise to help select and oversee the contractors. Although ARI is a research organization, it might be possible to arrange for ARI researchers to participate in the selection of contractors.

The next sections of this chapter addresses the supporting functions listed in Figure 2 in more detail. As noted, some of these functions are also discussed further in subsequent chapters.

Collect/Update Job Analysis Information

Test content, whether for a core competency assessment or MOS-specific assessments, should be determined based on a systematic analysis of job requirements. This, as with the other requirements we specify in this report, is intended to ensure the Army's testing program meets professional quality standards. Job analysis is the foundation by which the Army will be able to design an assessment process for each test (core and/or MOS-specific) that accurately and fairly reflects job requirements.

More specifically, the assessment program needs information about job requirements to (a) determine what content should be covered by the test, (b) what test method(s) would be most effective for assessing job competence within the bounds of cost constraints, (c) serve as the foundation for test questions, and (d) ensure the tests stay current. As discussed further in a later section on test development, there are many test methods other than multiple-choice and hands-on and many variations within each type. Job analysis ultimately results in test specifications that provide a recipe for the assessment.

Test Specifications

Test specifications indicate what types of assessments will be developed and exactly what they will look like. For a multiple-choice exam, for example, the test specifications should include a "blueprint" that specifies (a) how many items will be on the test, (b) what content areas will be covered on the test, and (c) how many items will be in each content area. Following are some considerations related to development of test specifications for an Army assessment program.

Tasks Versus Knowledges and Skills

The Army has traditionally defined job requirements in terms of very specific job tasks. To be successful in the long run, however, the tests used in an annual competency assessment program should be based on somewhat more broadly defined requirements. This will be invaluable in helping to ensure that test specifications do not shift significantly from one year to the next. Such fluctuations are undesirable for several reasons. First, it makes it difficult for Soldiers preparing for the tests because what they need to study changes from one year to the next. Second, it makes the program less effective because Soldiers will study just for exactly what is going to be on the test (e.g., particular tasks) rather than preparing in a broader fashion. Third, the program will be somewhat easier to maintain because it will be easier to develop and maintain item banks. A very important consideration is that using broader content areas makes it easier to accommodate differences in assignment-specific tasks performed by Soldiers in different locations/units. Looking at more broadly defined tasks and the knowledge and skills required to perform those tasks (e.g., knowledge of certain underlying principles) will make it easier to assess Soldiers doing similar activities with different pieces of equipment.

In addition to identifying what Soldiers do on the job and the knowledge and skills they need to perform their assignments, it would be best practice to characterize jobs in other ways that could help ensure that the resulting test specifications would be most accurate. This includes considerations such as contextual factors (e.g., being able to perform with frequent disruptions) and the extent to which critical job tasks require knowledge versus skill in manipulating objects that would ideally be reflected in a high quality assessment.

For the prototype tests we have developed during the course of the PerformM21 research effort, we did not collect detailed job analysis information from a broad sample of job experts as would be expected in the operational program. Thus, for example, the test blueprint for the prototype core competency assessment we developed with input from the ATPAT is not adequate for the operational program. Moreover, the prototype core assessment blueprint is largely task-based. As noted above, we recommend the Army adopt a somewhat less detailed focus for specifying test content, since detailed tasks change frequently. We developed principle- and systems-based test specifications for some of the MOS work described in Chapter 5, but again, this was based on the input of a fairly limited number of job experts.

Other Potential Test Content

Earlier guidance directed that the prototype core competency examination include sections on Army history and values. These are areas that are not likely to be supported in a typical occupational analysis effort. That is, if a survey asks Soldiers and NCOs the extent to which knowledge about various aspects of history impacts how well they perform their work each day, the answer is likely to be: "not very much." The Army can certainly determine that certain areas should be covered by the examination even if a clear job performance linkage cannot be supported. The other Services take this approach in their testing content and these areas are typically covered during the promotion board appearance, so there is established precedence for such a decision. We caution, however, that inclusion of these two particular areas presents challenges. One concern is Soldier reactions. It is important that Soldiers perceive the assessments to be relevant to their day-to-day jobs. Second, it is next to impossible to develop test questions on historical events that sound job relevant. This is in contrast to recommended test development procedures that specify questions be posed in performance-oriented contexts so they do not sound like they came out of a textbook. With Army values, it will likely be easier to establish an empirical link to job requirements, but it is difficult to develop test content that is challenging. It is pretty easy to figure out the answer to values-based questions and being able to answer the questions says little about the degree to which Soldiers behave consistently with those values.

A Summary of What is Needed

To meet the needs of a new assessment program, the Army needs a system for collecting job information with the following characteristics:

- Up-to-date information on job requirements collected with input from a broad sampling of job experts (preferably via survey and SME workshops)

- Job requirements specified using descriptors that are not equipment or assignment specific as is the situation with many traditional Army tasks
- Incorporates a process for determining what test formats will be used so additional supporting job information can be collected as needed
- Incorporates procedures for collecting supplemental information needed to develop certain types of assessments (e.g., complex simulations, situational judgment tests)

These requirements will ensure job relatedness of the tests. This will, in turn, help ensure the credibility of the assessment program with Soldiers and the ultimate success and effectiveness of the program.

Once the job analysis process is completed the first time in a manner that supports competency testing, it should be much easier to periodically update the information as needed. It is likely that we will generate recommendations for determining when new occupational analysis information should be collected to help ensure assessments stay current with field requirements.

For those MOS for which there are civilian certification programs that cover relevant job requirements, it may be possible to share resources. Certification programs typically conduct and document occupational analysis studies that form the basis for their examinations. Using this information, when it is relevant and accessible, has a couple of benefits. One would be to make it easier to develop MOS job descriptors that would be suitable for testing (e.g., competencies or principles and systems) and the other would be make the overlap in job requirements very explicit. This would facilitate transition assistance for veterans leaving the Army and make it more likely that civilian certification programs would give credit for related Army experience.

The Occupational Data, Analysis, Requirements, and Structure (ODARS) Program

The Army has in place an occupational analysis system that involves the periodic collection of data on job tasks and associated knowledge and skills (see Army Regulation 370-50). This is known as the Occupational Data, Analysis, Requirements, and Structure (ODARS) Program. The ARI Occupational Analysis Office provides support, but responsibility for conducting ODARS activities rests, under the TRADOC umbrella, with the MOS proponents. As discussed in the Phase I needs analysis report, currently existing job analysis information is insufficient to support the new assessment program for several reasons (R.C. Campbell & Knapp, 2004):

- Occupational analysis information for many MOS is out-of-date (i.e., more than 3 years old)
- The occupational analysis information is of uneven quality (e.g., with regard to task clarity, comprehensiveness, and overlap; adequacy of Soldier samples)
- The process as specified by AR 370-50 focuses on detailed job tasks
- Certain assessment methods that are likely to be adopted will require types of information not currently provided (e.g., critical incidents, walk-through analysis)

The first two problems listed above are largely resource issues that have been exacerbated by the high level of deployment activities over the last several years. Training products (e.g., course curricula, Soldiers' manuals) – the intended beneficiary of the ODARS Program – are suffering as a result (GAO, 2003). Ideally, the Army could adapt the ODARS to meet the needs of the new testing program so it does not have to embark on a totally new or largely redundant system. If this approach is to work, however, problems with the existing ODARS program would need to be addressed, in addition to incorporating new features into the program. If the Army addresses these needs by modifying the existing ODARS program, it will be important to ensure the requirements for supporting development of training resources and curricula are not compromised in the process.

One issue with the current ODARS Program is that the Army relies heavily on Army personnel (military and civilian) to carry out the job analysis work. Although TRADOC guidance specifies involvement of a training developer with experience in job analysis, it is unclear the extent to which most occupational analyses are implemented with such support. Although guidance is provided through Army regulations, there is no expert oversight to help ensure sufficient quality of the resulting work. Moreover, as suggested previously, it is not easy to do this type of work well. Like anything else, it is best to involve those who have been specifically trained for the requirement – providing written guidelines (as in TRADOC pamphlets) is not sufficient preparation in and of itself. We therefore recommend that, although proponents should retain responsibility for and support the collection of occupational analysis information, the work be conducted with the support of professional job analysts. This support could be provided through ARI (there is historical precedence for this, but additional resources would be required at ARI to support this strategy) or obtained through contractor support.

Prepare/Update Training and Doctrine

Even a testing program with a primary focus as a promotion tool will have a significant impact on training, training materials, and training delivery systems. Self-development programs must be in place that match and support the test effort, both for the Army-wide generalized Soldier competency areas and for any MOS-specific evaluation that eventually becomes part of the system. Test developers, trainers, and Soldiers must all work from the same resources and references and accessibility of materials is a major consideration for all. While the ATPAT indicated that the test system should not generate new, specialized training materials, up-to-date training materials must be available to support Soldier training and would, in turn, also support test design and development.

One area of critical interface is the Non-Commissioned Officer Education System (NCOES) and in particular the Primary Leadership Development Course (PLDC) and the Basic Non-Commissioned Officer Course (BNCOC). PLDC is a standardized, 4-week, common core resident course taught at various NCO Academies (NCOA), both in CONUS and OCONUS. PLDC is a prerequisite for promotion to Sergeant (E5). BNCOC is MOS-specific, varies in length by the MOS, and is normally a resident course at the MOS proponent location. All BNCOCs have a common-core component. BNCOC is prerequisite for promotion to Staff Sergeant (E6).

A test system and NCOES share too many common interests and threads not to be closely integrated. Both have the shared goal of preparing and qualifying the Soldier for promotion. As the testing program is developed and matured, the goal should be to maximize integration of the two programs, both doctrinally and administratively. This would be beneficial to both programs and could eventually lead to cost savings in the administration of the testing program.

Although Army training systems, per se, are outside the scope of our discussion, it bears emphasizing that training and assessment systems need to be mutually supporting. If training and doctrine resources (e.g., Soldiers' manuals) are not up-to-date and consistent with practice in the field, it will be much harder for test developers to construct high quality tests and it will be difficult for Soldiers to prepare for the tests. Indeed, it simply would not be fair to Soldiers to require them to take tests that do not reflect current job demands or for which adequate learning resources are unavailable. Moreover, there is some reason to believe that the need for up-to-date training materials for testing purposes might make performing those updates a higher priority for the Army.

A further organizational implication and consideration focuses on the alignment of the assessment function with the occupational analysis function. Operationally, the two requirements are closely intertwined, although occupational and job analysis also plays a role in other training and job definition needs as well. In the other services, the Air Force combines the two functions within the same organization, even physically co-locating them. The Navy employs a more remote, although symbiotic, relationship. The introduction of a measurement function should cause the Army to reexamine its occupational and job analysis function as well, including evaluation of the organizational structure for both. The assessment function will become a primary user of the products of the job analysis; if not organizationally joined, the two functions must have a clearly established operational relationship.

Design, Develop, and Update Assessments

Assessment Methods

There is much more to testing than traditional task-based multiple-choice questions and the hands-on testing that helped make the original SQT program so state-of-the art in its time but too resource-intensive to maintain. Alternative assessment methods are listed in Table 1. Each method has advantages and disadvantages and there are countless variations of each. For example, multiple-choice tests have a strong advantage over other test methods in that they can cover many job requirements relatively inexpensively. Even the most sophisticated multiple-choice tests, however, cannot evaluate certain job requirements (e.g., the psychomotor skill required to successfully shoot a target). Typically, the most fair and useful assessments include multiple methods to triangulate on the question of whether an individual has the required job knowledge and skill.

Some of the measurement methods listed in Table 3 are not going to be useful very often (e.g., on-the-job monitoring), but are likely to be quite desirable in selected situations. A key recommendation is that the Army a range of alternatives before selecting what is best in each situation.

Table 3. Assessment Methods

Performance-Oriented Multiple-Choice

- ✓ Questions posed in applied work contexts
- ✓ Visual aids to reduce reading and enhance realism (e.g., photos, figures)
- ✓ Animation to enhance realism
- ✓ Non-traditional item formats (e.g., matching, drag-and-drop)

Situational Judgment Tests

- ✓ Real-life problem scenarios depicted in writing or through video
- ✓ Examinees evaluate effectiveness of various possible actions
- ✓ Focus is on judgment rather than knowledge, per se
- ✓ Scoring key based on expert judgment (e.g., senior NCOs)

Path Simulation

- ✓ Examinees are presented with a computer simulation of a problem scenario
- ✓ Examinees progress through the simulation, stopping at various points to answer questions

Open Simulation

- ✓ Examinees are presented with a computer simulation of a problem scenario
- ✓ Examinees interact with the simulation, affecting how the scenario unfolds

Hands-On

- ✓ Examinees perform job tasks in a standardized environment
- ✓ Performance is scored by expert observers

Evaluation of Actual Work Products

- ✓ Samples of work products (e.g., recordings of a band members' performances) are scored by experts

On-the-Job Monitoring

- ✓ A process for recording performance indicators is embedded into operational systems
-

In the development of prototype MOS tests described in Chapter 5, PerformM21 project staff explored several variations on an array of assessment methods, including performance-oriented multiple-choice, situational judgment tests, path simulations, complex simulations, and hands-on tests. Lessons learned through this work will be discussed further in Chapter 5.

Once the assessment is conceptualized (in terms of form and content), it must be constructed with the input of job experts. In the Army's new assessment program, it will be necessary to have multiple forms of each assessment because it will not be practical to have all Soldiers take the exams at one time on a single day (as is the practice for the Navy). Fortunately, it is possible to construct computerized "banks" of test items that can facilitate development of alternate forms.

A Word About the Army's Current Approach to Testing

At this point, we should recognize TRADOC Pamphlet 350-70-5 (Updated August 2004), which is the guidance document for TRADOC organizations in how to apply testing within the context of Army training programs. Although this is a useful and comprehensive document, it is not entirely applicable to testing for purposes of assessment (as opposed to training). Note also that the document refers to two types of tests (see page 18, paragraph 3-4), neither of which

accurately describe the proposed competency assessment tests. The applicable text is reprinted in Figure 3. The Army assessment tests would be similar to criterion-referenced tests, in the sense that we recommend they be designed to be content valid. This means that the tests should cover a representative sample of job requirements regardless of how “discriminating” the content. For example, questions in some areas might be ones that most everyone gets right or most everyone gets wrong (hence they do not discriminate well among test-takers), but the content area would be retained because it is an important job requirement. Unlike the Army’s definition of a criterion-referenced test, however, we do not recommend establishing a performance standard (i.e., passing score) on each test. Rather Soldiers will receive promotion points based on how they score on the tests. Finally, although the tests are not designed to be norm-referenced, per se, scores will be reported that help Soldiers see how they have performed relative to other Soldiers and promotion decisions in the semi-centralized promotion process are based on a top-down ranking of promotion points.

a. The two major types/categories of tests are –

- (1) Criterion-referenced tests, which determine if learners can perform to established, well-defined training standards, or criterion (CRT are performance and knowledge-based tests).
- (2) Norm-referenced tests compare a learner’s performance with the performance of other learners (or the norm).

Figure 3. Criterion- and norm-referenced tests, as defined by TRADOC Pamphlet 350-70-5 (p.18).

Recommendations on Process

In the Phase I needs analysis report, Knapp and R.C. Campbell (2004) outlined the basic process for developing and maintaining content valid assessments. Here, we review the process (summarized in Figure 4), making observations and recommendations about how they would work in the Army assessment program. The emphasis will be on development of multiple-choice assessments, as these are likely to be at least part of any assessments developed within this testing program.

Underlying all process recommendations is a concern for test security. It is important that all the people and systems with access to confidential information (e.g., test questions, Soldiers’ test scores) handle that information in ways that preserve the integrity of the information.

Step 1. Draft, Review, and Revise Test Items/Materials

There are many guidelines associated with the development of high quality items (e.g., framing the question in terms of what is required in a job situation, avoiding “all of the above” options). Professional test developers facilitate the process by training SMEs to write effective items, then reviewing and editing the items as a quality control measure. This is a skill that requires both training and experience – it is hard for an SME to prepare high quality items the first time around. Test development contractors will often hire SMEs and thoroughly train them

on how to develop good questions. Work of these SMEs is then monitored by professional test developers.

We have developed a draft set of item writing guidelines suitable for the new Army promotion tests. These guidelines will help ensure high quality test items that will have the same “look and feel” regardless of who is developing them. It will also be necessary for the Army to determine what technology will be used to support the authoring of test items. There are commercial test authoring software tools available or the Army could develop its own system. In any event, the authoring tool should also easily link to (or be part of) a good item bank management software system. These requirements will be discussed further in Chapter 4 of this report.

Development

- Step 1. Draft, review, and revise test items/materials
- Step 2. Develop scoring protocols
- Step 4. Field test items/materials
- Step 5. Create test forms

Maintenance

- Add item statistics to item/test bank
- Review item/test bank for currency

Figure 4. Major test development and maintenance activities.

Development of high quality test questions is an iterative process. Draft questions should be subjected to review and revision by multiple SMEs. This helps to ensure that the resulting questions are clear and applicable to Soldiers in different settings. Ideally, the process will include small groups of SMEs (say three at a time) reviewing the draft questions together. The discussion that ensues provides a richer analysis of each question than that typically achieved by reviewers looking at each question independently.

The initial item development process is typically completed with the collection of “content validity” ratings for each new test item. The content validity ratings should be made by an independent group of SMEs who rate each item with regard to its relevance to the job, criticality (to avoid items that cover trivial information), and pertinence to the test blueprint/specifications. If the Army elected to use contractors for test development, an ideal scenario might be to have a group of Active duty personnel make the content validity judgments as a quality check on the contractor-generated test material.

There are significant variations on the development process required to support other less traditional assessment methods. For situational judgment tests, for example, content for new test items is usually generated through workshops with job incumbents. They generate “critical incidents” which are examples of actual job situations that then get translated into test questions. Hands-on tests that are highly proceduralized (i.e., there is only one approved way to perform the task) are relatively straightforward to develop and the Army is quite experienced with this method. For assessment (as opposed to training) purposes, however, more attention must be paid

to standardizing administration procedures and conditions and scoring protocols. Development of complex computerized simulations requires considerable time with individual SMEs to define the underlying variables in situations to be simulated and ways the computer program should respond to actions by the examinee.

Step 2. Develop Scoring Protocols

Step 2 is quite straightforward for traditional multiple-choice questions. Item developers should be required to identify a source (e.g., technical manual) that justifies the selection of a particular response option as the correct choice. In the item review and revision process other SMEs review the draft items and revise them to make them better. A particular area for concern in reviews of draft test items is the possibility of multiple correct answers.

Scoring other types of questions/assessments can be quite a bit more complicated. For example, scoring questions that are not strictly multiple-choice (e.g., matching questions) needs to be done in a way that does not weight the content of those items too heavily in relation to the 1-point multiple-choice questions. The process of developing a scoring protocol for situational judgment tests requires access to SMEs to judge the effectiveness of each response option for each new item. Scoring protocols for complex simulations can be even more complicated and must be done in a way that results in the fairest and most accurate representation of the examinee's ability to perform.

Step 3. Field Test Items/Materials

No matter how thorough the SME review and revision process is, unanticipated problems arise when the items are administered to real examinees. In this step, items are administered to examinees and the resulting item statistics are used to determine their final status – ready for operational use, need to revise, need to delete from the bank. Ideally, the field test of new items is embedded with administration of the operational assessment. For example, a 100-item exam might routinely include an additional 20 field test items (randomly distributed throughout the test) that are not used as a basis for calculating scores. Neither the Air Force or Navy explicitly field tests items, but they exclude clearly flawed items from the final test scores on each exam. That is, the final score on a 100-item exam may be based on just 95 items because five items may have had unacceptably poor item statistics.

Step 4. Create Test Forms

Previous steps in the process will result in a bank of usable test items. Prior to each test administration window, new test forms will be created from that bank. For multiple-choice exams, this involves drawing items from the item bank that conform to the exam blueprint. Each new form will have items that have been used before (on the immediately preceding form or in previous years) and newly field tested items. By testing all eligible personnel on the same day, the Navy simplifies things because they only require a single test form. Moreover everyone is tested every year, so there is no need to be able to compare scores on a test given one year to

scores on a test given a year later³. However, this is not a practical solution for today's Army. Therefore, multiple test forms will be developed for each assessment period. If scores are to be valid across multiple assessment windows (which is required unless every Soldier is tested each year), the forms also need to be equivalent across assessment periods.

We recommend creating multiple test forms using a psychometric method (item response theory [IRT]) that allows creation of equivalent test forms with a wide combination of items. We recommend linear (as opposed to adaptive) testing to help ensure the content validity of each test form. Once the test forms are generated, they should be reviewed by at least a couple of job experts to make sure the test items are current and that the items are non-overlapping. As the item bank grows larger during the life of the assessment program, it will become increasingly difficult to ensure every item in it is current and certainly there will be items that are too similar to each other to use on the same test form.

After these tests have been administered and before scores are finalized, the data should be analyzed to conduct a variety of quality control checks. These analyses will include estimates of test form reliability and checks for potential test compromise.

At this point, there is no generally accepted method for generating equivalent forms for certain types of assessments (e.g., situational judgment tests). We are exploring some of the relevant considerations as part of this project.

Figure 4 does not include a step to set a cut score for each assessment that is required for establishing pass/fail cutoffs. Since the Army intends to use test scores as a basis for awarding promotion points, it will not be necessary to set a pass score on the assessments. This is good because establishing pass scores is a very difficult process that is frequently done incorrectly. It is much better not to have to do it. We recommend awarding promotion points in a manner that requires a simple statistical transformation of test scores (that might range from 0 to 200, for example) to promotion board points (that might range, for example, from 0-150).

Maintenance

Test questions should be managed in an automated "item bank." An item bank typically includes information such as the following:

- Item text and correct answer
- When written
- Status (ready to be field tested, ready for operational use, requires revision)
- Item statistics (classical item statistics and IRT parameter estimates)
- Reference that documents correct answer

³ Test items vary in difficulty, so it would not be fair to randomly select items from an item bank to create a unique test for each Soldier. That is, some Soldiers would get a harder test than others unless one statistically alters the scores on the different forms to make them comparable.

Items do not necessarily remain static once they are entered into an “item bank.” Rather, it is important to update information about the questions (e.g., with updated item statistics) and revise questions as needed to keep them current.

Scheduling, Records Management, and Communication with Examinees

For purposes of this discussion, we are referring here to activities that take place in preparation for test administration. These activities include the following:

- Determining which Soldiers are eligible for testing
- Providing Soldiers with information they need to prepare for testing
- Scheduling Soldiers for testing
- Answering Soldiers questions

If only because of their large number, it is no small requirement to communicate with the Soldiers who will be taking assessments each year. Telling Soldiers about the assessments, scheduling them for testing, providing score reports, and answering questions is a major undertaking.

During the prototype core pilot test, we tried to mimic some of the communication requirements using Soldiers’ Army Knowledge Online (AKO) email accounts. As described in Chapter 3, this strategy was generally not successful. It appears that many Soldiers simply do not use the AKO system. While this would likely change if they knew this would be the only way to receive important information (e.g., about promotion test dates), it is worth noting that the security features associated with AKO (e.g., complex password requirements, needing to frequently change passwords) make it frustrating for users.

General Information

Basic information about the assessment program should be available on a web site devoted to the program. This information would presumably include the following:

- Test dates and locations
- How the assessments are developed
- How the assessments are scored and what type of feedback is provided to examinees
- Quality control measures to ensure tests are fair and accurately scored
- Frequently asked questions

In addition to a web site, the Navy conducts traveling “briefs to the fleet” to inform enlisted personnel about their assessment program. Varying the way in which information is presented (in-person, paper-based materials, web-based materials, text-based presentations, oral presentations, graphic presentations) helps to ensure everyone has access and really understands the message. It also helps to keep the message simple. It is not uncommon to make certain program design decisions (e.g., how scores will be calculated) in part on how simple the explanation of the process can be made for examinees.

Misinformation and misunderstandings about an assessment program can be very damaging. It is also true that it will not be possible to completely prevent this from happening and that it is human nature to disparage tests that one has to take. Nonetheless, there is much the Army can and should do to promote open and honest communications with Soldiers about the assessment program.

Soldier Test Preparation Guides

Soldiers should be given a guide to help them prepare for the assessment test. In civilian certification programs, a test preparation or study guide typically includes information about the following areas:

- What is the purpose of the test
- What is on the test (e.g., a test outline or blueprint)
- How many items are on the test and what they look like (often sample items are included)
- A bibliography of sources that would be useful for studying in preparation for the test
- What the test experience will be like (e.g., test locations, time limits)
- Hints for taking the test (e.g., there is no penalty for guessing so do not skip any questions)
- Warnings about cheating and the consequences
- How the test will be scored
- How scores will be used
- What type of feedback will be provided and when it will be provided

The test preparation guide may be completely automated or it may refer to a web site for further information or a sample test. Particularly when an assessment is provided in a novel format (as at least used to be the case with computer-based testing), a sample test that familiarizes the examinees with the test administration format prior to test day is highly desirable. It is important that test preparation guides be accurate and provided to examinees in sufficient time to allow them to prepare for the examination.

Although outside the scope of the assessment program, per se, a prototype self-assessment program has been developed. The initial version is targeted to knowledge and skills covered by the core assessment. In the self-assessment exercise, Soldiers get feedback on their responses to individual test questions and self-assessment scores do not become part of their official records. The proposed operational self-assessment program would have a broader focus – helping Soldiers gauge their promotion potential with regard to all elements included on the Promotion Point Worksheet. To effectively serve Soldiers' needs, both the test preparation guides and self-assessment exercises must be backed up with up-to-date training material and job aids.

Deliver Assessments

By the time the Army's competency assessment program is fully implemented, it will involve delivering assessments to tens of thousands of enlisted Soldiers each year. The advantages of delivering tests on computers are fairly obvious. Managing and maintaining

security while shipping paper tests to and from sites all over the world is a monumental, time-consuming task. Tests must be ready to go long before the scheduled administration period and cannot be changed once they have been printed in volume. Loss of even one test booklet compromises the entire process. Therefore, we assume that as much of the test delivery as possible needs to be automated.

Challenges

The test delivery system needs to reach Soldiers in the Active and Reserve forces. It needs to reach Soldiers who are deployed or the assessment system will not work for purposes of promotion testing.

The biggest challenge to computer-based testing in a large-scale program is having an infrastructure for delivery – that is, secure facilities set up with suitable computers and monitored by test proctors. In any high stakes assessment program, examinees will be motivated to score as well as they can. It would certainly be possible to deliver tests over the Internet to Soldiers just about anywhere, but it would compromise the integrity of the program to do so. The Army needs to know that the right Soldier is taking the assessment and that the Soldier has no unfair advantage over other examinees (e.g., in terms of prior knowledge about what will be on the test or access to resource materials during the test). This is an assessment program, not a training exercise.

Available Resources

Ironically, it is the infrastructure the Army has set up to support training – the Distributed Learning System (DLS) with its supporting Digital Training Facilities (DTFs) and Distributive Training Technology Project (DTTPs) – that offers the most promise for handling most of the test delivery needs associated with the new competency assessment program. These facilities are well-equipped, but currently unevenly utilized. Although the DLS network and organizational orientation is decidedly training centered, precedence exists for utilization of this resource to deliver testing. An ARI initiative supporting Fort Jackson recruiter training qualification uses DTFs to administer Internet testing of the NCO Leader Skills Inventory (NLSI).

Matters related to test delivery are discussed in considerably more detail in Chapter 4. Information, observations, and recommendations in that chapter reflect experience we obtained pilot testing the Army-wide core competency prototype examination and updated information regarding the Army's computer facilities and next generation Learning Management System.

Scoring and Score Reporting

At the end of the test window for a given examination, it will be necessary to analyze test data to ensure the accuracy of results before scores are disseminated. The post-test analyses were described briefly in an earlier section. This section focuses on score reporting, which has three components:

- Ensuring test scores are integrated into Soldiers' personnel records
- Providing test results to Soldiers
- Providing summary results to Army leaders

Integrating Scores into Personnel Records

This is primarily a technology issue and, as such, will be discussed further in Chapter 4 of this report. It goes without saying that the transfer of test scores needs to be accomplished efficiently, accurately, and confidentially.

Soldier Feedback Reports

Typically, examinees would like as much feedback as possible on an assessment, including which individual items they answered incorrectly. Providing such detail is not feasible in most operational testing programs because the test items get re-used on future test forms. It generally is possible, however, to provide an indication of how well the examinee performed in various areas covered by the assessment. The key is making sure scores provided to examinees are statistically reliable, so it would be fine to report a subscore based on 20 test items but a subscore based on four items would not be informative.

Examinees intuitively understand “number correct” scores (e.g., “I got 65 out of 100 items right”). Unfortunately, once we have multiple forms of an exam that are likely to vary somewhat in difficulty level, it is necessary to statistically transform scores from the different forms to make them comparable to each other. The good news is that this is possible and it is commonly done (e.g., with college entrance exams). The process, however, can be difficult to communicate to examinees.

Examinees also want to know how they performed relative to others. Therefore, we envision a Soldier feedback report that includes the following elements:

- Standardized scores – total and for each major component of the test)
- Norm-referenced scores (e.g., percentile scores) – total and for each major component of the test
- Bar graphs to depict the profile of strengths and weaknesses as indicated by the subtest scores and to show performance relative to the Soldiers peers
- Text descriptions of how to interpret the information provided and where to go with questions

As described further in Chapter 3, we developed a feedback form to provide to those Soldiers who participated in the Army-wide core assessment pilot test.

Summary Results for Army Leaders

It is reasonable to expect that Army leaders will be interested in seeing how well Soldiers in general are performing on these promotion tests. Determining exactly what information to provide and to whom will be an important exercise. The key is to provide enough information to be useful while avoiding the potential for misuse of the information. An example of misuse would include comparison of units’ performance on composite test results or the evaluation of commanders based directly on how their unit “scored.”

One way this process might work is to develop a summary report that is analogous to the Soldier Feedback Report, but which includes average total scores and subscores for various groups of interest. The groups might be defined by component, major command, or type of MOS (in the case of the core competency exam).

Program Evaluation

As the Army's new assessment program is implemented, problems will inevitably surface and require modifications to program procedures and policies. It would be to the Army's advantage to proactively look for these problems and areas that might be improved. Through program evaluation, problems can be identified and addressed before damage occurs. We recommend that evaluation strategies be incorporated into all phases of program implementation and become a permanent part of the program. Not only would this be in the Army's best interests, but it would also address provisions of the Government Performance and Results Act (GPRA) as recommended in a recent GAO review of TRADOC activities (GAO, 2003).

We also recommend managing the expectations of program stakeholders (e.g., Soldiers, proponents, test developers) so they appreciate the willingness of and need for the Army to change the program to be successful over time. It will not be a static entity that, once in place, will not change.

Program evaluation should start even before operational implementation. It is anticipated that there will be a trial period or ramp-up prior to implementation. Although initial test scores would not be used for promotion decisions, this nonetheless would provide an opportunity for accruing baseline performance data to be used for confirmatory analysis in an operational program.

Evaluation Methods

Program evaluation can take many forms. Researchers typically speak of formative evaluation and summative evaluation strategies. Formative evaluation generally involves obtaining stakeholder reactions to a program (e.g., examinees, test developers, test delivery personnel). It is usually fairly easy to incorporate formative evaluation processes as an ongoing feature of a program (e.g., to routinely solicit feedback on test items from examinees whenever an exam is given). Summative evaluation is more demanding, in that the goal is to provide empirical evidence that the program is meeting its intended goals. For example, it might require a research project in which assessment test scores are correlated with subsequent job performance at the next pay grade.

Concerns that should be explored in program evaluation efforts include the following:

- To what extent are the assessments reliable and content valid?
- Are unnecessary costs (in terms of money and time) being incurred?
- Are support organizations, including contractors, performing adequately?
- Are there inefficiencies in how the program operates across MOS and supporting offices?
- Are the Soldiers who perform best on the assessments the same Soldiers who perform the best on the job?
- Is the assessment program promoting more effective training and self-development?

Managing Expectations

Closely related to the idea of program evaluation is the issue of stakeholder expectations. This is particularly true in testing programs because those who have already been tested tend to resist changes to the program that might be viewed as making it easier for future examinees. Thus, they expect that once the assessment program is in place, it will become the standard and will not change. Those responsible for running the assessment program may also resist change because it requires that they change how they administer the program. This type of reaction ties the hands of those interested in improving the program as lessons are learned and new technology offers new solutions. Thus, it is important to portray the assessment program as one that will evolve as needed to reflect sound testing practice and to promote efficiency of operations.

Summary

It should be clear from the preceding discussion of the functions that will need to be performed to support a new Army competency assessment program that such a program will be a major undertaking for the Army. Successful implementation will require change, invention, and adaptability. It is crucial that, from the start, the assessment program not be viewed as a stand-alone or separate ancillary program. It must be fully integrated with other existing programs.

The Army has been without any type of formalized Army wide testing system since the early 1990s; probably 90% of the current force has had no first hand experience with an Army testing system. Start-up and implementation of a test system will have a profound effect on many aspects of the Army command and culture. Foremost, the testing system must receive recognition and support from the highest command levels if it is to become viable, supported, and accepted by the junior enlisted and those within the NCO ranks who it will impact the most. It is significant that the impetus for the revival of the Army testing program came from the ranks of the NCOs – the 35,000 Soldiers who participated in the ATLDP were quite explicit in their description of what is required. Endorsement of their findings and a continuing commitment to implementation by the Chief of Staff of the Army (CSA) will be the first of several required senior Army leadership interventions needed to facilitate the program.

Key to acceptance by the Army is understanding the system and key to this understanding is knowledge. Soldiers must perceive testing as fair and equitable and all aspects of testing must be open and transparent. Long before the system is implemented there needs to be a program that explains and publicizes the system. Lack of coherent and consistent information leads to confusion and the substitution of rumor and legend in place of fact. The Army needs to take advantage of existing Public Affairs, Strategic Communications, and other troop information channels to publicize all aspects of the program. Moreover, the Army needs to establish a centralized testing website where Soldiers can go not only to find information but to also ask questions and seek clarification and receive personalized, real-time feedback. Finally the Army needs to aggressively push the program through spokespersons that travel to the field and explain the program, face to face, with Soldiers. Erroneous beliefs were a major factor in the demise of the SQT but the fact that they were erroneous is immaterial – the important factor is that they were believed. A modern program, based on solid, stable policies, communicated openly using multiple strategies and taking maximum advantage of email, websites, and other technology, is an essential piece of the Army test system.

CHAPTER 3: PILOT TEST OF THE ARMY-WIDE EXAMINATION

Karen O. Moriarty, Tonia Heffner, Roy C. Campbell, Huy Le, and Deirdre J. Knapp⁴

Background

Development of a prototype core examination began at the initiation of the PerformM21 research program. The purpose of developing one prototype test prior to the others was to provide insights into an operational testing process while simultaneously identifying the theoretical and technological advances in testing since the demise of the SQT. The entire process from test development to administration was intended to simulate an operational test. It is one thing to speculate about anticipated issues or challenges, but quite another to actually face and handle them. In addition to learning about the testing process from Soldier notification to test administration to giving Soldier feedback, all of which was to occur electronically, we also wanted to create a bank of usable job knowledge items. Although we did not initially intend to develop a prototype core assessment to investigate these issues, the SMA's guidance identified a core test as the most practical choice.

In Phase I, we relied on SMA and ATPAT guidance to develop a test blueprint as opposed to job analysis, which is recommended for an operational assessment. The prototype blueprint specified a 150-point assessment, and based on the specifications, HumRRO staff wrote new items and adapted items from previous research projects (i.e., Select21 [Knapp, 2003] and Project A [J. P. Campbell & Knapp, 2001]). SMEs from the Army and HumRRO reviewed all items. At the end of the Phase I item development and review process, we had an item bank with 282 test items (R.C. Campbell, Keenan, Moriarty, Knapp, & Heffner, 2004). In addition to the job knowledge items, we decided to administer a situational judgment test (known as the "LeadEx") to assess leadership. The LeadEx was developed and validated in the NCO21 project, described in Chapter 1, for potential inclusion in the semi-centralized promotion system (Ford, R. C. Campbell, J. P. Campbell, Knapp, & Walker, 2000; Knapp et al., 2002). At the conclusion of Phase I, we were prepared to begin pilot testing the core examination.

Method

Data Collection Plans and Issues

The ATPAT members recommended the pilot test be delivered via the Internet using DTF/DTTP facilities, and use of these facilities, we were told, required prior registration of the Soldiers with the Army's Distributed Learning System (DLS). Registration with DLS was necessary because the test was to be authenticated through the DLS central server, which would generate a unique ID that it would pass to Questionmark'sTM server, which was hosting the assessment. This ID would then allow test data to be connected with identification information in Army Training Requirements and Resources System (ATRRS). This server interaction would provide security to both the test content and Soldier personal information in compliance with the Privacy Act. Because the pilot test results were confidential, instead of providing the Army with

⁴ Considerable support for the work described here in this chapter was provided by Sonia Kim, Jennifer Solberg, Joshua Updegraff, and Shonna Waters.

Soldier results, the server interaction process was only going to indicate which Soldiers took part in the pilot test. However, there were data transfer issues that could not be resolved in time to use this process in the pilot test.

The original plan specified that the project team would receive from the tasked unit a roster of Soldiers (including names and AKO email addresses) participating in the pilot test approximately 2 months before the test date. In turn, we would send an email notification to the Soldiers advising them of the upcoming pilot test and their requirement to register with the local DTF/DTP prior to the test date. This email would (a) instruct the Soldiers where and how to register with the DTFs, (b) request that they contact a member of the project staff to confirm their participation, (c) include a test preparation guide, and (d) encourage them to seriously prepare for the pilot test. The *DCAP Test Preparation Guide* was a document created to familiarize the Soldiers with the project, test content areas, item types, and study references (see Appendix A).

The original plan was altered because of unanticipated practical realities. Some of these changes were welcomed while others created unique challenges.

- The pre-registration requirement was dropped when we learned that project staff would be able to register Soldiers with the information provided on the roster or use generic registrations. This change was welcome because of the anticipated problems with Soldiers failing to register in time, not to mention the hardship this may have created for Reserve and National Guard Soldiers who may not live in close proximity to their assigned posts.
- The tasking process used by most units did not allow for a roster to be available 2 months prior to testing. Complete rosters were received only for three sites (see Table 4), which were all Reserve units.
- Soldiers were typically informed the day of the test that they were participating. When asked on the post-test survey about when they were notified, approximately 82% of the Soldiers noted they learned of their participation within the last week. That was the response option with the shortest time interval. The lack of an advance roster meant that Soldiers did not receive the *DCAP Test Preparation Guide* and therefore, were not able to prepare. Only the 99th RRC and Fort Dix groups were sent the test preparation guide.
- We learned from the 99th RRC and Fort Dix groups that most of the Soldiers who did receive the email notification did not check their AKO email accounts often, if at all. The few who did receive and read the email did not look at the attachment (the *DCAP Test Preparation Guide*). Similarly, many Active Soldiers in the other pilot test sessions did not know their AKO email accounts' passwords. Even if we had been able to send them email notifications, many would not have read them.

Data Collection

Sample Description

We collected data from 571 Soldiers from 6 Active Component and 3 Reserve Component locations (see Table 4). The sample was 76% male, 58% white, 21% black; 20% reported being of Hispanic origin. There were 90 MOS represented, with the largest percentages coming from 11B (11%), 19K (7%), 92A (7%), 68W (6%), and 92F (6%). Eighty-seven percent were in the E4 paygrade, and mean time in service (TIS) and time in grade (TIG) were 43.56 months and 17.41 months, respectively.

Approximately 45-50 Soldiers participated in focus group sessions immediately following the test administration. These were informal meetings sometimes held as a group, and sometimes one-on-one. As Soldiers completed the test, we would ask them if they would agree to participate in an informal discussion of the testing experience. We tried to balance the participants in terms of race, gender, and time to complete the test.

Table 4. Core Exam Pilot Test Locations

Date	Location	Number of Soldiers	Roster Received	Reserve or Active Component	Test Version
3 April 2004	99 th RRC	19	Yes	Reserve	1
1 May 2004	Fort Dix	20	Yes	Reserve	1
7 – 11 June 2004	Fort Riley	91	No	Active	1
21 – 28 June 2004	Camp Humphreys	64	No	Active	1
19 – 23 July 2004	Fort Lewis	133	Partial	Active	1
2 – 6 August 2004	Fort Bragg	44	No	Active	1
11 September 2004	653rd Area Support Group	54	Yes	Reserve	2
12– 15 October 2004	Fort Hood	46	No	Active	2
November 2004	Schofield Barracks	100	No	Active	2
	TOTAL	571			

Test Administration

In each pilot test session, four assessments were presented: the Soldier Background Form, the job knowledge items, the situational judgment test (LeadEx), and the Soldier Reaction Survey. The background form collected typical demographic data such as race, gender, and MOS along with “recency” and “frequency” ratings involving use of, or experience with, various weapons (e.g., M9, M240B) or competency areas (e.g., First Aid, Conducting Drill and Ceremony). The test items (i.e., job knowledge and situational judgment) were presented in four different orders to discourage cheating and to ensure we collected enough data on all items (see Figure 5).

The total job knowledge item bank contained 282 items, with some of these worth more than 1 point. The goal was a 150-point job knowledge item test. However, we also wanted to gather data on as many of the 282 items as possible. We determined the best way to achieve this goal was to administer a subset of approximately 190-200 items (called Version 1) to a couple

hundred Soldiers and run item analyses to select the best set of items yielding 150 points. These items would form the core item set. Then we would administer this core set plus another 40 to 50 items to a couple hundred more Soldiers, calling this Version 2. Table 4 shows which locations received Version 1 and which received Version 2. The number in each cell represents the number of Soldiers tested at each location.

Following administration of the computerized tests, the Soldier Reaction Survey was administered to get information from Soldiers regarding their reactions to the test process and the *DCAP Test Preparation Guide*. The form asked for Soldiers' feedback on areas such as:

- Computer-based testing
- Using the DTF for such a test
- Their test performance
- The perceived appropriateness of the pilot test items

We also conducted informal focus group and one-on-one discussions with a subset of Soldier participants to obtain additional reactions and ideas related to the prototype assessments.

Aside from a number of technological issues, which are mentioned below and discussed more thoroughly in Chapter 4, the actual test administration went smoothly. There was at least one project staff member present to review the project briefing and the Privacy Act Statement, help with computer problems, answer questions, and discourage cheating. After the test, the administrator conducted the focus groups or interviews with Soldiers.

ORDER A	ORDER C
Soldier Background Form	Soldier Background Form
Core Items	Reverse Order Core Items
LeadEX Items	LeadEx Items
Soldier Reaction Survey	Soldier Reaction Survey
ORDER B	ORDER D
Soldier Background Form	Soldier Background Form
LeadEx Items	LeadEx Items
Core Items	Reverse Order Core Items
Soldier Reaction Survey	Soldier Reaction Survey

Figure 5. Presentation order.

Recall that this pilot program had two goals: (a) test development and (b) determining the operational implications of implementing a testing program. We collected data to make recommendations regarding both goals. For the most part, the test data informed test development, while Soldier reactions and project staff observations informed operational implications.

We decided at the outset of this project to provide Soldiers with feedback on their performance so we could get their reactions to a prototype feedback report. During the administration process we collected email addresses for participating Soldiers so we could later email them their feedback. The feedback provided to Soldiers was based on the core items selected during the analyses and not intended to be very detailed. Essentially, we wanted to let

Soldiers know how they did on the core items and how they compared to the others taking part in the pilot test. Appendix B is an example of the feedback provided to Soldiers.

Technical Problems

All but one administration occurred in a DTF. The administration at the 99th RRC in Pittsburgh occurred in a leased computer facility. Many technical problems were experienced with test administration. As shown in Table 5, the technical issues encountered can be grouped into three categories: DTF login problems, computer-specific problems, and system-wide problems. Recall that Soldiers were required to login to the DLS prior to beginning the pilot test. The manner in which this process was handled varied by location. Fort Dix created personal logins for each Soldier based on the roster. However, the roster contained many errors such as misspelled names or incorrect email addresses. This, in turn, caused problems as Soldiers attempted to login using the correct spelling of their names. Camp Humphreys and Forts Lewis, Bragg, and Hood used generic logins so that all Soldiers logged in using the same login. Fort Lewis also used generic logins, but there were errors in the process such that some Soldiers could not login. Schofield Barracks' DTF manager set up the computers ahead of time so that the Soldiers did not have to login to the DLS.

The locations varied in terms of the resources of the computers and resulting computer-specific problems. For instance, the drag and drop items required a Java applet to function properly. Most machines had this applet, but some did not. Many computers had trouble loading graphics. The job knowledge items are grouped into the following five sections: Common Tasks Skill Level 1, Common Tasks Skill Level 2, History/Values, Leadership, and Training. The number of items in each section ranged from 30 to 95, each section loading as one big web page. Some machines had the resources to load all items and associated graphics correctly and some did not. Beginning with the 653rd ASG pilot, we changed the presentation of the items so that the largest section contained 62 items. Thereafter, there were no more similar problems.

Table 5. Summary of Technological Issues

Location	DTF Login Problems	Computer Specific Problems	System-wide Crashes or Problems*
99 th RRC	No – not a DTF	No	Yes
Fort Dix	Yes – incomplete roster with errors	No	No
Fort Riley	No – used a generic login	Yes	Yes
Camp Humphreys	No – used a generic login	Yes	Yes
Fort Lewis	Yes – generic logins did not work	Yes	Yes
Fort Bragg	No – used a generic login	Yes	Yes
653 rd Area Support Group	No	No	No
Fort Hood	No	No	Yes
Schofield Barracks	No – computers already set-up by DTF manager such that no login required	No	Yes

* Refers to multiple computer crashes.

System-wide problems included situations where multiple machines in a location would repeatedly freeze, crash, or time out. Often Soldiers would complete one section of the test, click the "Continue" or "Submit all Answers" button to advance to the next page or section, and get

bumped off the system. This was sometimes solved by simply having Soldiers log back in and start up where they left off. However, sometimes when they logged back in the Soldiers were returned to the beginning of their most recent section and were not able to determine if their data were saved or not. In most cases the data were saved, but the Soldiers, not knowing this, often completed the section a second time.

Soldier Reactions

Focus Group Responses

Most of the feedback received in the focus groups or interviews was quite positive. This was true whether it was Soldiers who finished quickly, or those who were unable to complete the test in the allotted time. Computer-based tests did not pose a problem for most of these Soldiers as many took the Armed Services Vocational Aptitude Battery (ASVAB) on a computer. The non-traditional items (e.g., matching, drag and drop) were a welcome change from multiple-choice. Many of the Soldiers commented that the test reminded them of all they had forgotten since basic training. Some of the areas that posed the most difficulty were Weapons, Land Navigation, and Leadership, whereas the Values section was perceived as the easiest.

The *DCAP Test Preparation Guide* was presented for their comments. They liked the idea of a preparation guide for such a test and thought it would be a useful guide. In particular, they liked the fact that the guide showed which topics would be covered and provided names of reference manuals. Many Soldiers requested and were given copies of the guide.

Soldier Reaction Survey Responses

Of the 571 Soldiers who participated in the pilot test, 520 completed the Soldier Reaction Survey. We asked questions to gauge their reactions to computer-based testing, testing in the DTF, the notification process, the test preparation process, and the effectiveness of the pilot test items. Approximately 76% have taken computer-based tests before, which confirms what they told us in the focus group sessions. When asked if they had a preference for computer-based tests over paper-and-pencil tests, approximately 11% responded "no," 65% responded "yes," and 25% indicated "no preference." Regarding taking tests in the DTF, only 11% reported having ever used a DTF prior to the pilot test. Seventy-six percent either agreed or strongly agreed that the DTF is a good place to administer a test like the pilot. These responses bode well for using computerized testing and DTFs in the future.

Soldiers were asked to gauge their performance on the pilot test. Specifically, they were asked, "How well do you think you did on the Common Task (Army/NCO History, Training, Army Values, or Leadership Items)?" Table 6 shows the results as percentages. Not surprisingly, the middle choice (neither well nor poorly) was the most selected in all content areas except the Army Values Item. Overall they seemed to think they did not do poorly, with about one third of them indicating they did "well" or "very well."

We also asked Soldiers to indicate the effectiveness of the items to measure their knowledge of the different content areas. Specifically they were asked, "Assume you had all the

time you needed to prepare for the pilot test. How effectively do you think the test would measure your knowledge of Common Tasks (Army/NCO History, Training, Army Values, or Leadership)?” Table 7 shows the results as percentages. The results here are positive. For each area, over 70% of the Soldiers indicated the items would measure their knowledge either “well” or “very well.”

Table 6. Soldiers’ Ratings of Their Performance

	Very Poorly	Poorly	Neither Well nor Poorly	Well	Very Well
Common Task Items	3.9	13.3	49.4	30.3	3.1
Army/NCO History Items	5.0	20.3	44.3	26.9	3.5
Leadership Items*	4.1	10.1	44.7	35.2	6.0
Training Items	4.3	13.2	46.2	32.9	3.5
Army Values Item**	2.3	6.6	32.7	42.7	15.7

Note. $n = 517-518$; numbers in cells are percentages.

*Excluding the situational judgment test (LeadEx) items.

**This was a single, matching item requiring the Soldiers to match the seven Army Values with their correct definitions.

Table 7. Soldiers’ Ratings of Item Effectiveness

	Very Poorly	Poorly	Neither Well nor Poorly	Well	Very Well
Common Task Items	1.5	4.6	22.9	43.7	27.2
Army/NCO History Items	1.7	5.8	20.7	42.7	29.0
Leadership Items*	1.9	3.7	24.1	44.6	25.7
Training Items	1.5	2.9	21.8	46.1	27.6
Army Values Item**	1.2	2.9	19.0	40.8	36.2

Note. $n = 517-519$; numbers in cells are percentages.

*Excluding the situational judgment test (LeadEx) items.

**This was a single, matching item requiring the Soldiers to match the seven Army Values with their correct definitions.

The survey results generally support the positive feedback we received during the focus groups. This reflects a belief by most of the Soldiers that a competency test of this nature has value. Perhaps they feel a competency assessment program would be able to identify and remedy performance issues before they become critical. However, whether the feedback will remain as positive during an operational test is unknown. The stress of an operational test may color some Soldiers’ responses as they shift their focus from competency assessment as an abstract program to how an operational test would affect them personally.

Core Job Knowledge Exam Item Analyses

A total of 266 items were pilot tested, of which 24 used non-traditional formats. Version 1 consisted of 192 items. Based on the item statistics, we created a 150-item core set that approximated the blueprint specified by the ATPAT. We created Version 2 by combining the core item set with 89 new items, but discovered that during pilot testing, due to a Questionmark™ technical error, 19 of the Version 2 items were not administered. This included

10 from the core item set. The steps taken to develop and analyze the core item set are discussed below.

Selecting the Core Item Set

Version 1 was administered to 371 Soldiers (see Table 4). However, analyses were run only for Soldiers completing all items, reducing our sample size to 156. To select what would form our core item set, we arranged the Version 1 items within their content areas and sub-areas by item-total correlation. Then, items were selected to simultaneously maximize overall average item-total correlation and coverage of the 150-point blueprint. Despite our best attempts, we did not match the blueprint exactly because there were some low-performing items that could not be included.

Of the 16 non-traditional items originally presented as part of test Version 1, 13 were selected to be part of the core item set based on the item-total correlations. This suggests that these items as a whole performed well. The non-traditional items were worth multiple points, so we needed to derive appropriate weights. For example, if a drag and drop item has five pieces that must be dragged to their correct locations, there are three ways to think about the “worth” of that item. First, the item could be worth 5 raw points – one for each piece that is correctly dragged and dropped. This may overweight the item relative to the traditional multiple-choice items. Second, the item could be worth 1 point – credit given only for correctly dragging and dropping all five pieces. This may under-value the item relative to a traditional item. Third, one could empirically determine a maximum weight for the item that reflects its informational value.

We opted for the third strategy. The method described below is one way to handle weighting non-traditional items. However, in an operational testing program with larger numbers of examinees, it would be possible to use simpler IRT-based procedures to accomplish the same end.

Estimating Weights for Nontraditional Items

Confirmatory factor analysis was used to estimate the reliabilities of the non-traditional items in order to determine the weights that maximize the composite score formed by those items and the multiple-choice items (Drewes, 2000; Wainer & Thissen, 2001). Five content area composites were formed by summing the selected traditional multiple-choice items belonging to the respective domains. The covariances between these composites and the non-traditional items were used as the input for this analysis. In the measurement model tested, the residual errors of the non-traditional items and the content domain composites were allowed to be correlated if they belonged to the same content domains. The model yielded very good fit ($\chi^2=110.32$, $df=106$, $p=.258$ CFI=.987, GFI=.961, RMSEA=.012, SRMR = .040), indicating the appropriateness of the theoretically expected structure of the test items.

Reliabilities of the non-traditional items were estimated by their squared multiple correlations (i.e., ratio of variance accounted for by the performance construct and respective domain-specificity). Formulas suggested by Wainer and Thissen (2001, formulas 25, 26, and 27, pp. 45-46) were used to estimate the optimal weights for the non-traditional items that would maximize total score reliability.

Table 8 shows the total raw points for the items along with the unit weights estimated by the procedure. Multiplying the unit weight for each non-traditional item by its raw points yields the “optimal points” contributed by the non-traditional item to the total score. These optimal points provide rough assessments for the “worthiness” of the non-traditional items in relation to that of the multiple-choice items (e.g., a non-traditional item having an optimal point value of 3.00 can be considered equivalent to three multiple-choice items). For ease of interpretation (and also to facilitate the feedback process), we eliminated the decimals by rounding the optimal points. As a result, weights for non-traditional items were adjusted accordingly.

Table 8. Estimated Weights for the Non-Traditional Items

Item Category	Total Raw Points	Unit Weights	Optimal Points	Rounded Points	Rounded Weights	Item Type
Skill Level 1 CT	4.000	0.345	1.382	1.000	0.250	Ranking
Skill Level 1 CT	3.000	0.431	1.294	1.000	0.333	Ranking
Skill Level 1 CT	5.000	0.620	3.102	3.000	0.600	Matrix
Skill Level 1 CT	5.000	0.797	3.985	4.000	0.800	Matching
Skill Level 2 CT	5.000	0.410	2.049	2.000	0.400	Matching
Skill Level 2 CT	2.000	1.109	2.218	2.000	1.000	Drag & Drop
Leadership	5.000	0.772	3.859	4.000	0.800	Matrix
Leadership	5.000	0.936	4.679	5.000	1.000	Matrix
Leadership	5.000	0.975	4.877	5.000	1.000	Matrix
Training	5.000	0.220	1.100	1.000	0.250	Matrix
Training	4.000	0.619	2.720	3.000	0.750	Matching
Training	5.000	0.730	3.649	4.000	0.800	Matching
MC Composite ^a	1.000	1.000	1.000	1.000	1.000	Multiple-Choice

Note. CT = Common Tasks; An example of a matrix item is one that presents a series of tasks and asks the Soldier to indicate for each task whether it is an NCO or Officer duty.

^a This composite was formed by summing scores of all the selected multiple-choice items (86 items).

Note that Table 8 contains only 12 non-traditional items. The Army Values matching item was the only values item and was worth 7 points. Because there are seven Army Values, we decided to retain this item with its raw points (i.e., 7) and not estimate a weight for it.

As shown in Table 8, the non-traditional items performed well with the exception of the ranking items, which tended to have lower item-total correlations. This resulted in generally lower estimates of their optimal weights. While further data are needed to corroborate the findings, it appears that developing other types of non-traditional items (i.e., matrix, matching, drag and drop) may be more efficient, especially when the number of items that can be included in the final tests is limited.

The reliability obtained for the total score by applying optimal weights for the non-traditional items is only minimally higher than that obtained based on a simple sum of item raw scores (.886 and .882, respectively). The unweighted reliability is based on the raw scores and the weighted reliability is based on the application of the optimal weights from Table 8. This minimal gain at the total score level is because coefficient alpha was already high due to the relatively large number of items included. It is difficult to obtain further substantial increases to

the internal consistency reliability. Further, weighting is arguably necessary because it provides a logical means for determining the number of points that a non-traditional item should contribute to the total test score (i.e., an item contribution to the total score is based on the "amount of information" the item has). The optimal points enable assessment of the value of each non-traditional item vis-à-vis the traditional multiple-choice items, which in turn permits selection and allocation of items during scale construction that maintain the distribution of test points specified by the test blueprint.

Table 9. Distribution of Points and Items in the Core Item Set

	Number of Multiple-Choice Items	Number of Non-Traditional Items	Core Item Set		Blueprint Specifications	
			Number of Raw Points*	Percent of the Test	Number of Points	Percent of the Test
Skill 1	45	4	54	42.18%	69	46.00%
Skill 2	12	2	16	12.50%	20	13.33%
History/Values	14	1	21	16.41%	23	15.33%
Leadership	7	3	21	16.41%	20	13.33%
Training	8	3	16	12.50%	18	12.00%
<i>Total</i>	<i>86</i>	<i>13</i>	<i>128</i>	<i>100.00%</i>	<i>150</i>	<i>100.00%</i>

*Raw points refer to the points prior to applying the non-traditional item optimal weights shown in Table 8.

Based on these results, a subset of 109 items worth a total of 138 raw points was selected to be the core item set. The core item set did not exactly match the blueprint specifications. The blueprint was a very specific, task-based blueprint, and we were not able to develop and pilot test as many items as we would have liked for all the categories. Therefore, at times, when an item performed poorly, we were not able to replace it with a better performing item. As a result of a Perception™ technical problem, we were forced to drop 10 core multiple-choice items, which were worth a total of 10 points. This outcome resulted in an even greater deviation from the blueprint because there was no item coverage for the following Skill Level 1 sub-areas: Combat Techniques (Survive), Navigate (Mounted and Dismounted), and Remains Reporting and Handling. Table 9 describes the final core item set and contrasts it to the blueprint specifications. Table 10 provides the summary item statistics for the core item set based on the full sample.

Table 10. Summary Item Statistics for Core Item Set based on Full Sample

	Mean Corrected Item-Total Correlation	Mean Item Difficulty	Weighted Reliability
Skill Level 1	.11	-.23	.75
Skill Level 2	.08	-.19	.39
History/Values	.16	-.09	.56
Leadership	.16	-.09	.53
Training	.13	-.21	.54
<i>Total Score</i>	<i>.12</i>	<i>-.23</i>	<i>.86</i>

Note. $n = 338$ after listwise deletion of full sample ($n = 571$). Weighted reliabilities were estimated based on a formula for reliability of weighted composite score (Feldt & Brennan, 1989; Wainer & Thissen, 1999).

Deriving Scores

Because this was a research project, Soldiers were not required to answer all (or any) of the items. This, and the technological issues described earlier, resulted in a considerable amount of missing data. In an operational program, items with missing data are treated as incorrect. Therefore, in an operational program Soldiers should be explicitly encouraged, if not required, to answer all items because it is to their advantage to do so. For this research effort, however, the following decision rules were developed to handle missing data when creating scale scores:

- If more than 10% of the data were missing for any one scale (e.g., Skill Level 1 or Leadership) then a score was not generated.
- If 10% or less of the data were missing, then the missing items were ignored and a scale score was derived based on the remaining items.

We deviated somewhat from the blueprint, so when deriving the total score, we weighted each subscale by its blueprint weight to match the blueprint specifications. For example, the blueprint specified that the Skill Level 1 content area should account for 46% of the test. Therefore, the Skill Level 1 scale was given a weight of .46. Likewise, weights of .1333 were applied to the Skill Level 2 and Leadership content areas, and weights of .1533 and .12 were applied to the History/Values and Training content areas, respectively.

The procedure for scoring many non-traditional items is very simple – count the correct responses to each item stimulus. However, for matching items where the number of stimuli equals the number of response options, scoring should range from 1 to $k-1$ (with k = number of stimuli/options) because of the dependence of the two last responses (cf. Budescu, 1988). The same logic can be extended to scoring for ranking items. As an example, one matching item with $k = 4$ was recoded to take this dependence into account (i.e., scoring it from 0 to 3), and this slightly improved the item-total correlation for this item from .078 to .082.

Descriptive Statistics of the Core Item Set

Descriptive statistics for the core item set and the subscales are shown in Table 11. Overall, Soldiers got an average of 60% of the items correct. When asked how well they did on the job knowledge items, the Soldiers tended to respond “neither well nor poorly” (see Table 6). Approximately one third of them believed they did “well” or “very well.” The results here suggest that they were a little high in some areas, but fairly close overall, in the assessment of their performance.

Table 11. Descriptive Statistics of Core Item Set

Score/Subscore	Total Number of Items	Mean Score	SD	Min Points	Max Points	Mean Percentage Correct
Skill Level 1	49	40.27	8.24	14.31	61.45	58%
Skill Level 2	14	10.59	2.62	2.50	16.50	53%
History / Values	15	15.46	3.64	2.19	23.00	67%
Leadership	10	11.17	2.55	2.59	18.50	56%
Training	11	11.48	2.76	1.97	17.72	64%
Total Score	99	90.20	15.45	39.85	132.73	60%

Note. The mean score may exceed the total number of items because some items are worth more than one point.

LeadEx Analysis

The LeadEx situational judgment test was developed for a related Army project (NCO21; Knapp et al., 2004). It measures the following constructs:

- Relating to and Supporting Peers
- Cultural Tolerance
- Motivating, Leading, and Supporting Individual Subordinates
- Training Others
- Directing, Monitoring, and Supervising Individual Subordinates
- Problem Solving/Decision Making Skill
- Team Leadership

Although developed to cover multiple constructs, the LeadEx provides a single overall score that reflects general leadership. The estimated reliability for this sample was .80, which is comparable to what was obtained in the NCO21 project. The estimated reliability reported for the NCO21 sample of E4 Soldiers was .76.

The core examination total score and LeadEx score were moderately correlated ($r = .45$). While one would expect the two assessments to be correlated, it is gratifying to see that they are not highly related. The finding suggests that together they provide a more comprehensive assessment of Soldiers' competence. One also might expect the LeadEx to correlate most highly with the Leadership subscore of the core job knowledge exam, however it was correlated fairly similarly with all of the areas except History/Values (see Table 12). This might reflect low amounts of construct overlap between the two leadership scores, with the core exam items tapping clearly knowledge-based content (e.g., Policies and Procedures of the Chain of Command) whereas the LeadEx focuses on leadership skills (e.g., problem solving). This finding is also consistent with the large literature on managerial assessment centers that shows that there are higher correlations between different constructs measured with the same method (i.e., in the same exercise) than between the same constructs measures with different methods (Lievens, 2002; Neidig & Neidig, 1984).

Table 12. Correlations among Job Knowledge and LeadEx Scores

Scores	1	2	3	4	5	6
1 Skill Level 1						
2 Skill Level 2	0.51					
3 History/ Values	0.46	0.39				
4 Leadership	0.47	0.45	0.48			
5 Training	0.48	0.43	0.47	0.52		
6 Total Score	0.89	0.69	0.71	0.70	0.69	
7 LeadEx	0.42	0.39	0.26	0.37	0.35	0.45

Note. $n = 334\text{--}467$. All correlations are statistically significant at $p < .01$.

Subgroup Differences

Table 13 contains the results of the subgroup difference analyses for the job knowledge items. Soldiers were allowed to indicate membership in more than one racial group. Therefore,

sample sizes differ based on which comparison is being made. Using Cohen's (1988) framework, these effect sizes are moderate to large, particularly for the White-Black and White-Hispanic comparisons. The White-Asian comparison differences were more moderate in size. These race/ethnicity results are what one would expect, based on the literature in high-stakes testing in employment and education (Sackett, Schmitt, Ellingson, & Kabin, 2001). Although we would expect both the job knowledge and LeadEx tests to exhibit race differences in an operational setting, these may be reduced when Soldiers have the opportunity to prepare for the tests.

Table 13. Subgroup Performance Differences on the Core Job Knowledge Item Set

Group	<i>n</i>	<i>M</i>	<i>SD</i>	Group	<i>n</i>	<i>M</i>	<i>SD</i>	<i>d</i>
White	211	94.84	14.14	Black	64	79.62	15.40	-1.08*
White	193	96.06	13.24	Hispanic	56	81.64	13.02	-1.09*
White	214	94.85	14.11	Asian	24	88.72	15.29	-0.43*
Male	259	92.34	15.36	Female	79	83.18	13.65	-0.60*

Note. Effect sizes were calculated as the differences between the means of the two groups divided by the standard deviation of the referent (i.e., White or male) group. * $p < .05$.

The LeadEx showed much smaller subgroup differences, which is surprising because the LeadEx requires more reading than the core items. Perhaps the reading requirement is offset by the nature of what is being assessed – that is, experience-based judgment. The LeadEx difference reported for the White-Black comparison is larger than that reported in the NCO21 report ($d = .32$; Knapp et al., 2004).

Table 14. Subgroup Performance Differences on the LeadEx Items

Group	<i>n</i>	<i>M</i>	<i>SD</i>	Group	<i>n</i>	<i>M</i>	<i>SD</i>	<i>d</i>
White	310	1.92	0.74	Black	114	1.58	0.87	-0.46*
White	278	1.95	0.73	Hispanic	104	1.62	0.76	-0.45*
White	313	1.92	0.74	Asian	41	1.71	0.78	-0.28
Male	413	1.80	0.79	Female	132	1.86	0.72	0.08

Note. Effect sizes were calculated as the differences between the means of the two groups divided by the standard deviation of the referent (i.e., White or male) group. * $p < .05$.

Female Soldiers performed moderately less well on the core examination and performed essentially equal to male Soldiers on the LeadEx (see the last rows of Tables 3.10 and 3.11). The result for the LeadEx is consistent with prior research reported for NCO21 (Knapp et al., 2004). The gender difference on the core examination may be due to the fact that 26% of our sample comes from combat arms MOS, which are closed to women and which are more likely to focus their training on the basic Soldiering skills.

We examined performance differences among the MOS categories of Combat Arms (CA), Combat Support (CS), and Combat Service Support (CSS) Soldiers. As Table 15 shows, the CSS group performed significantly lower than the other two groups on the core Job Knowledge item set. Perhaps this is because CA and CS Soldiers spend more of their time practicing and training on general Soldiering skills. CS Soldiers performed significantly better than their CA and CSS counterparts on the LeadEx.

Table 15. MOS Category Differences on the Core Job Knowledge Item Set and LeadEx

Scale/Sub-scale	Total Group					Combat Arms (CA)		Combat Support (CS)		Combat Service Support (CSS)	
	<i>M</i>	<i>SD</i>	<i>D</i> ₁₂	<i>D</i> ₁₃	<i>D</i> ₂₃	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Job Knowledge	90.20	15.45	0.37	-0.47	-1.03	93.53	14.68	98.97	12.04	86.58	15.52
LeadEx	1.82	0.77	0.39	0.00	-0.49	1.75	0.79	2.06	0.63	1.75	0.79

Note. $n_{CA}=81-137$, $n_{CS}=52-93$, $n_{CSS}=187-283$, d_{12} = Effect size for CS-CA mean difference, d_{13} = Effect size for CSS-CA mean difference. d_{23} = Effect size for CSS-CS mean difference Effect sizes calculated as (mean of non-referent group-mean of referent group)/*SD* of referent group. Referent groups are listed second in the effect size subscript. Statistically significant effect sizes are bolded, $p < .05$ (two-tailed).

Computer-based testing is growing in popularity. As noted before, approximately 76% of the Soldiers told us during the pilot test that they had taken a computer-based test (CBT) before. Surprisingly, there were significant, moderate differences in test performance between those who reported having taken a CBT before ($n = 251$; $M = 92.32$; $SD = 14.74$) and those who reported not having taken one before ($n = 75$; $M = 84.00$; $SD = 14.83$). The effect size ($d = -.54$) was significant at $p < .05$. It could be that being unfamiliar with CBT affected performance by making Soldiers too anxious about, or distracted by, the process to pay close attention to the test itself. With the growing popularity of CBT, it is possible that differences will fade over time. We examined the correlations between CBT and test performance to see if age moderated this relationship. Partialling out age did not reduce the magnitude of the correlations.

Table 16 shows correlations between some demographic variables and test performance. Both age and time in service were correlated with the LeadEx. The LeadEx measures skills that are expected to improve as the Soldier becomes more experienced in the Army, so this is not surprising. Time in service also correlated with the Skill Level 2 scale, which is also not surprising since those with more time in service would have more opportunity to be exposed to higher level task requirements. Finally, experience with weapons correlated relatively highly with the Skill Level 1 scale, probably because weapons-related items comprised 47% of that scale. Experience with weapons was also correlated with the Skill Level 2 scale, but not to the same magnitude.

Table 16. Correlations Between Demographic Variables and Test Performance

	Age	Time in Service	Time in Grade	Experience with Weapons
Skill Level 1	.00	.09	.04	.27*
Skill Level 2	.05	.11*	.07	.10*
History / Values	.00	.01	.03	.06
Leadership	.00	.03	.03	.08
Training	-.02	.05	.06	.04
Total Score	.03	.09	.05	.21*
LeadEx	.11*	.10*	.06	.02

* $p < .01$

Providing Feedback to Soldiers

We asked Soldiers to provide us with their email address so we could send them feedback on their pilot test performance. This feedback included numeric and graphic descriptions of the

Soldier's scores on each of the subsections as well as the total score. Of the 571 Soldiers who participated in the pilot test, 523 provided us with email addresses and were subsequently sent their feedback. Some Soldiers provided us with more than one email address (e.g., their AKO account and a civilian account), and in that case we sent the feedback report to all addresses provided. This meant that a total of 786 emails were sent, and of those, 129 (16%) were returned. Emails were returned primarily due to spelling, "no such account," or inactive account errors. We received only five responses from Soldiers. Two were from Soldiers asking a specific question about the feedback report, two were from Soldiers advising us they were not the intended recipient, and one was an automatic "out of office" reply. This suggests that the feedback report was fairly easy to understand and that this method of providing feedback has promise for an operational test. Actually, in an operational test, the number of returns may be substantially lower because the Army would have access to Soldiers' correct AKO email account addresses, avoiding many of the spelling errors we encountered.

As mentioned earlier, Soldiers were not required to answer the items, which resulted in a considerable amount of missing data. While the analyses just described were run using listwise deletion, feedback was provided to as many Soldiers as possible. Missing scale scores were explained on the feedback provided to Soldiers with a note that there were too many missing items to generate a score.

Pilot Test Products

There are a number of products resulting from the pilot test efforts. These are listed and briefly discussed below:

- Prototype test form
- Item bank with statistics
- Prototype Test Preparation Guide (Appendix A)
- Prototype Test Administrator's Manual
- Prototype Feedback Report (Appendix B)
- *Item Development Guide* (Appendix C)

The prototype test is comprised of 128 points (99 items). The total item bank includes approximately 282 items, of which 266 were administered at a minimum of one location and include statistics such as item total correlation and item difficulty. The prototype test preparation guide was intended to help Soldiers prepare for the pilot test. However, it will also make a good template for an operational test preparation guide. The same can be said for the prototype test administrator's manual and prototype feedback report. The *Item Development Guide* can be incorporated into item developer training for an operational test.

Recommendations for an Operational Test

The recommendations below are meant to supplement those discussed in the previous chapter. Please refer to Chapter 2 for a more complete listing of recommendations for an Army operational test.

Blueprint of the Future

The DCAP blueprint specifies the tasks and other areas to be covered at a very detailed level. We recommend an operational blueprint allow for broader sampling of potential test content by using categories of tasks or other elements. An operational testing program will require multiple equivalent test forms, and for some tasks there are only a certain number of items that can be developed. Unless the Army wants to test the same tasks each year, the blueprint would have to change each time a new set of test forms was developed. Even if the Army was willing to test the same tasks each year, the procedures associated with those tasks would evolve. Frequently changing the blueprint is resource-intensive for both item development and creating and keeping Soldier test preparation guides current.

A demonstration will clarify the different approaches. Suppose the information in Table 17 provides the top-down importance rankings for the tasks from the category, First Aid. If First Aid was to account for 12 points of a 150-point test, then the top three or four tasks would be selected and items would be generated for only those tasks to allow for multiple items for each task. However, for a category-focused blueprint, the top 8 to 10 tasks would be selected and used as the basis for the category definition. For example:

First Aid: Test items cover topics such as providing emergency care or treatment to fellow Soldiers who are injured in training or combat. Items encompass tasks that range from evaluating the extent of injury to treating typical combat injuries in order to prevent worsening of the situation or death.

Items would be developed for these tasks, but for each administration items totaling 12 points would be sampled from a large pool.

Table 17. Top-Down Rankings of First Aid Tasks

Mean Importance Ranking	SD	First Aid Task
3.25	4.81	Evaluate a casualty
4.83	4.09	Perform Mouth-to-Mouth Resuscitation
5.33	2.74	Perform First Aid for Bleeding of an Extremity
6.67	3.89	Perform First Aid to Clear an Object Stuck in the Throat of a Conscious Casualty
6.67	3.92	Perform First Aid for Heat Injuries
6.92	3.80	Perform First Aid for an Open Abdominal Wound
7.25	2.67	Perform First Aid for an Open Head Wound
7.25	3.57	Perform First Aid for an Open Chest Wound
7.83	3.61	Perform First Aid to Prevent or Control Shock
9.25	2.30	Perform First Aid for Burns
9.92	3.48	Perform First Aid for a Suspected Fracture
10.00	5.17	Practice Individual Preventive Medicine Countermeasures
10.75	3.31	Perform First Aid for Cold Injuries
10.75	3.39	Perform First Aid for Nerve Agent Injury
13.25	2.80	Transport a Casualty

An even more drastic change from Army precedence is developing test blueprints that are not task-based at all. As discussed in Chapter 5, we developed MOS-specific test questions targeting principles and systems rather than job tasks. For example, a mechanic's job can be thought of as a series of tasks (e.g., replace master cylinder, correct malfunctioning batteries), or a constellation of competencies (e.g., knowledge of hydraulics, knowledge of basic electrical principles). A major advantage of this approach is that it is easier to specify test content that is generalizable across different settings (e.g., different pieces of equipment). It is also generally easier to write questions geared to a knowledge-based blueprint than to one that is task-based.

The benefits of using a category-based blueprint (whether or not it is oriented to tasks) are not only in terms of resources needed to create and maintain the item bank and study resources. There are also benefits for Soldier development. Telling Soldiers exactly which tasks or detailed topics will be covered will lead them to focus on those specific areas to the exclusion of others in the category. Telling Soldiers what types or categories of material will be covered will encourage them to prepare more broadly. This difference is very significant. It makes "training to the test" a positive, rather than a negative, outcome.

Including Skill Level 2 Content on a Skill Level 1 Test

Most job knowledge and skill tests include only current level content. Skill level 2 content was included on the pilot test because the ATPAT felt this was a better way to gauge proficiency for performance at the E5 level. While collecting the MOS job analysis data, a process that is discussed more in depth in Chapter 5, we first tried to conceptualize appropriate test content in terms of "E4 Soldiers eligible for promotion." This was a difficult judgment for the NCOs serving as our subject matter experts at our first site visit. They believed that an "E4 Soldier eligible for promotion" should actually be performing at the level of an E5 Soldier, not an E4 Soldier. Based on this feedback we then tried "E4 Soldiers with 3 years time in service" at our subsequent site visits. This also proved to be a difficult judgment for our SMEs, most of whom felt that if Soldiers are going to be tested on skill level 2 content then they should receive training on those tasks.

Our item analyses for the DCAP, discussed in more depth above, support these sentiments. Table 18 summarizes the item analyses results for each content domain. Skill Level 2 items did not perform as well as the other items. The Skill Level 2 content domain had the lowest reliability and the lowest item total correlation. Leadership and Training each had fewer items, but higher coefficient alphas and average item total correlations.

Table 18. Summary Item Statistics for Each Content Domain

Content Area	Number of Items	Average Corrected Item-Total Correlation	Weighted Reliabilities
Skill Level 1	49	.11	.75
Skill Level 2	14	.08	.39
History/Values	15	.16	.56
Leadership	10	.16	.53
Training	11	.13	.54

Our recommendation is not to include content from the next skill level on an operational test. However, this is ultimately a decision that should be based on policy factoring in job analysis survey data.

Use of Non-Traditional Items in an Operational Setting

Non-traditional items can provide a unique way to gather performance data for Soldiers. Care must be taken when creating these items, however, to ensure maximum psychometric value. For instance, in matching items, where possible, the item writer should provide more response options than stimuli. Item writing guidelines for traditional, multiple-choice items are easily available. However, this is not the case for non-traditional items. We developed an item writing guideline that includes information about traditional and non-traditional items and included it in Appendix C.

As mentioned, non-traditional items provide a unique way to measure Soldier performance. Drag and drop, animated graphics, and even simulations are possible with much of today's item development software, including Questionmark's Perception™. However, these items come with costs. First, these items take longer to administer than traditional multiple-choice items. For instance, for the second version of the pilot test, the traditional multiple-choice items took on average approximately 22 seconds to complete compared to 52 seconds for the non-traditional items. For a test of 100 items this can add considerable time. None of these non-traditional items included animated graphics or simulations, which would most likely take even more time to complete. The MOS pilot tests will include such non-traditional items. Second, non-traditional items cost more to develop. Third, many non-traditional items are more complex to score, requiring additional analyses to determine how they should be weighted.

It is unknown whether these items provide additional information concerning Soldier performance over their traditional counterparts. If they do, does this information justify the increase in administration time and costs? Obviously, the best way to determine this would be to collect criterion data and compare the relationships of the different item types with this data. However, short of this, there are some questions that developers can consider to help them decide which item type is best. For instance, when deciding whether to include a graphic, the item developer might want to consider the following:

- Does the graphic serve a *real* purpose? Including graphics simply because that capability exists is not advisable.
- Does the graphic reduce the text and, thus, the amount of reading required?
- Is there a suitable graphic available, or will the item developer need to create or modify one?
- To be effective, does the graphic need to be animated?

Similarly, when developing a matching item, the developer will want to ensure that this is the best way to present the item. A matching item is actually a combination of several multiple-choice items, and should be more efficient in terms of amount of reading and space required. Other concerns include whether the different stimuli and response options actually belong together in one item, and whether there are any formatting issues for those stimuli and response options with the software.

Item Presentation (One-at-a-Time vs. Grouped)

For the first version of the pilot test, the items were presented in the following groups: Common Tasks Skill Level 1, Common Tasks Skills Level 2, History/Values, Training, and Leadership. Within each group, Soldiers could navigate to any item by using the Previous Question or Next Question buttons, or using the scroll bar at the bottom of the screen (see Figure 6). As mentioned in the discussion of computer-specific problems, this means that all the items in a group are downloaded as a single web page, which can strain computer resources. In the first version of the pilot, the largest group was Common Tasks Skill Level 1, which contained 95 items. For the second version of the pilot test, we split this group into a weapons group and general task group, which meant that the largest group would contain only 62 items. Reducing the web page size eliminated the computer-specific problems in the last three pilot test administrations.

There are two factors that affect the decision to present computerized items grouped or not: (a) technological (or resource) issues and (b) testing philosophy. The paragraph above touches on one of the technological issues: presenting items in groups is taxing on a single computer's resources. However, presenting items in groups is also taxing on computer *systems*. In Perception™, when items are presented one-at-a-time, as the test-taker advances to a new item the data for the old item are automatically saved. When items are presented as a group, the data are automatically saved when the test-taker advances to the next *group* of items. With a large group of items, it is wise to incorporate another data saving mechanism that saves data periodically (i.e., every 3 or 5 minutes) while test-takers are within an item group. However, this can be taxing on a computer system (e.g., a DTF) if multiple computers within the system are trying to write data to the test server at once. This is because periodic, timed saves occur at the same interval (i.e., every 3 or 5 minutes), whereas saving after each item or each group of items is driven by test-takers' progress, which is likely to differ from person to person.

To some extent with the pilot test, this may have been alleviated by Soldiers taking the assessments in different orders. However, it is possible that during a pilot test administration at least half of the computers were trying to initiate a "save" at about the same time because this is driven by elapsed time rather than by Soldier progress, which is likely to be more variable. It is possible that this caused *some* of the system-wide issues. It could not account for all of the system-wide problems because we instituted the intermediate save procedure in response to the large number of system-wide crashes we experienced at the first several data collections.

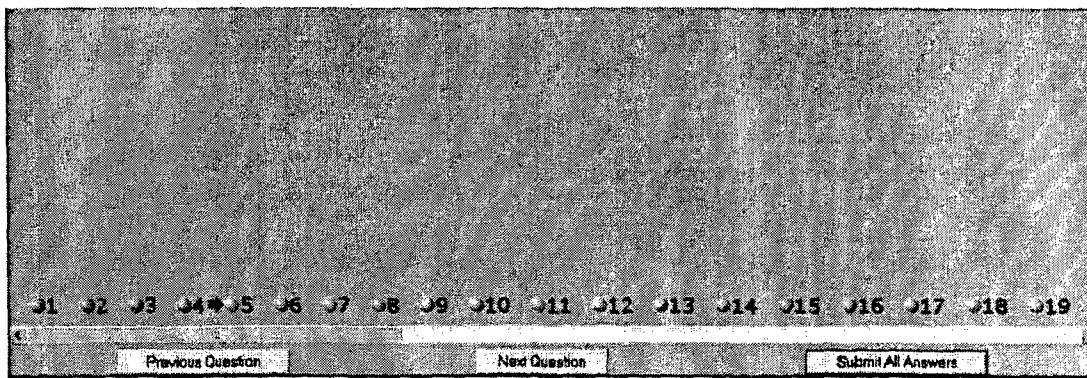


Figure 6. Screenshot of navigation tools for Soldiers taking the core examination.

Testing philosophy also influences the decision about how to present items. Some test developers believe that test-takers ought to be able to go back to previous items and change their answers just as they can with paper-and-pencil tests. Some organizations may include that as part of their testing policy. The reason the core exam items were presented in groups was to allow Soldiers to revisit previous items to make changes should they desire. It was impractical to present all pilot test items as one group, so the decision was made to present them in several groups based on their content.

Presenting items all at once is a security threat. It allows Soldiers, at the completion of a test, to scan all items in an attempt to commit as many as possible to memory in order to write them down upon exiting the test room. This phenomenon is called item "piracy" and is a common problem in high stakes testing programs. Presenting the items in groups reduces this risk and presenting items one-at-a-time reduces the risk further.

Multiple Forms

It would not be possible for all Soldiers to take the same test at the same time. Testing most likely will be completed in windows (e.g., 3-4 months). The longer the test window, the more exposure test items would have, and this translates into the need to develop more items to support more forms.

Simply randomly selecting items from a bank to create new tests for each administration does not guarantee that Soldiers are taking equivalent tests. Item response theory (IRT) can be used to develop equivalent forms, which would be necessary to ensure scores are valid across multiple assessment windows.

Test Preparation and Administration

In an operational test environment, prior Soldier notification would be required. The amount of time Soldiers would need to prepare for such a test is debatable, but when asked how much preparation time would be needed, 2 weeks was the median and modal response from the pilot test Soldiers. They may have significant misconceptions about what is necessary to prepare for such an important test. We recommend that Soldiers be notified of the test requirement and provided with test preparation guidance at least 4 to 6 months prior to test administration. However, as the program is rolled out, information about testing cycles will be routinely available to all Soldiers so that they could begin their own preparation at any time.

A test preparation guide similar to the one created for this assessment would be needed. If possible, a test preparation website should be created that includes links to electronic versions of the required study manuals, sample items, a frequently asked questions (FAQ) section, information about the test content areas, and study tips. However, even though a test website or preparation guide is available for Soldiers, the bigger issue is whether Soldiers will have the time and resources with which to prepare.

The configuration of the DTF workstations is another tool to counter unacceptable test behaviors. For instance, currently in the DTFs computer desks are constructed in such a way as to prevent Soldiers from easily accessing ports on the back of CPUs. The front plates of CPUs

and peripheral keyboards do not house USB or other ports to which data storage devices might be attached. As the Army systems are upgraded, this should remain the case. Media storage devices and related equipment such as digital cameras are becoming smaller and more sophisticated, making cheating somewhat easier. For any operational test, it is recommended that the Army use proctors and remain vigilant for potential methods Soldiers may use to compromise test security.

The technological issues will need to be remedied prior to any operational test being launched. High stakes testing is stressful without all the stops and starts required as a result of such problems, not to mention the resulting confusion over whether or not the responses were properly saved. Further, these are legitimate reasons for Soldiers to contest the results. Prior to making this test operational, Army DLS/DTF personnel will need to discuss in detail technological requirements with item and software developers. These issues are discussed further in Chapter 4.

CHAPTER 4: TECHNOLOGY ISSUES AND OPTIONS

Roy C. Campbell, Jeffrey A. Barnes, and Shelly West

Introduction

Perhaps no other single feature of the contemplated Army assessment system is more critical than the issue of the technology supporting and surrounding the system. Even predating the initiation of the PerformM21 project, the assumption was that “technology” would be the driving factor in enabling a revised Army testing program. The initial guidance of the Sergeant Major of the Army, which established his vision of an Army assessment program, relied heavily on “technology” with references to “computer-based” and “web-administered” as distinguishing characteristics. The review of the Army’s past experiences with testing – the SQT system – conducted as part of this project (Knapp & R. C. Campbell, 2004) identified inefficiencies in many aspects of that program – test development, distribution, administration, scoring and analysis, notification – that collectively contributed to the demise of that otherwise effective system. Many of these inefficiencies, while endemic to a manual-based system as required in the SQT era, could conceivably be eliminated through applications of technology. By way of contrast, the SQT required large numbers of people to support even logistical and routine facets; a 21st century assessment system would be built around computers and networks, allowing people to concentrate in those areas requiring human intervention, decisions, and input.

But embracing “technology” as a concept is much different than implementing a technology-based system. The realities of incorporating technology into assessment, while still expectant, pose many challenges and are, in many ways, unique to specialized fields in the overall area of technology. However, it is critical that the technology aspect of the Army assessment research not be isolated or viewed as independent of the other design themes and that all principals – including Army decision makers – have a basic understanding of technology capabilities and of the issues that must be faced.

During the first year of the project (Phase I) we took a preliminary measure of the status of what we perceived as the technology issues (Knapp & R. C. Campbell, 2004). The areas that we included then have proven to be enduring, however, we have learned much through actual experiences with technology during test development and the pilot testing of the core test during Phase II. Additionally, we have developed a more robust model of the possible testing requirements and have had an opportunity to gather more current information on the status of technology. This chapter then, both summarizes and updates original information and offers insights and emerging requirements gathered through empirical observations.

General Issues

This chapter addresses technology applications in three major aspects of the system: Test Development, Test Delivery, and Administrative Support. Each of these will be covered in detail in this chapter. However, there are several broad issues that apply to all facets of the system.

System Definition

What, exactly, does technology need to support? Chapter 2 of this report defines the current projection of what an Army assessment program could look like. While it is an evolving picture, based on research goals and done prior to any implementation decisions, it nonetheless provides a working premise. The program is multi-faceted, involving both Army-wide core testing and eventually MOS test development spread across proponents. Additionally, it includes Soldier notification, Soldier self-assessment and preparation, record keeping and reporting, and test maintenance and sustainment.

Likewise, “technology” is a broad area. It includes hardware and software but also communication and data links and other access issues. It also includes databases, website maintenance and management, system security, broadband issues, and systems integration. While the technology will rely predominately on commercially available off the shelf technology systems (COTS), there will undoubtedly be generated some custom requirements and engineering, particularly in software/courseware support.

Figure 7 is a prototype depiction of the system requirements. Most of the emphasis in this interpretation is on the testing process. Chapter 2 contains a narrative description of the test development requirements and that function is not detailed here. However, it need be noted that test development could eventually involve linked systems at multiple proponent locations. This process overview is fundamental to system understanding and technology anticipation. It should be kept in mind during the ensuing discussion in this chapter of individual functions and applications. Figure 7 will be addressed again in more detail in the discussion of the Administrative Support requirements at the close of this chapter.

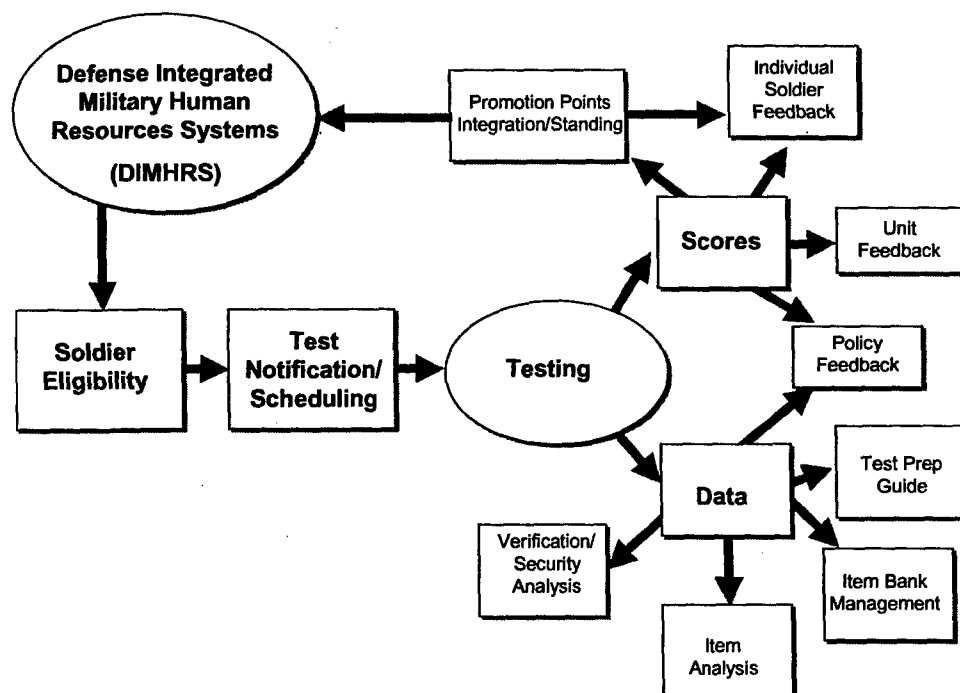


Figure 7. Test administration requirements model.

Planning and Integration

Basic to the concept of testing technology is an intrinsic assumption that the testing system will be built around existing or planned Army systems; that is, there will not be a separate or independent "testing network." As depicted in Figure 7, testing will be integrated with the promotion function of the emerging Defense Integrated Military Human Resources System (DIMHRS) personnel operations operational environment. It will rely on existing networks such as AKO for Soldier notification, feedback, and for access to test preparation and self-assessment tools. Testing will likely draw on hardware and management systems such as found in the Digital Training Facilities (DTFs) for test administration as described later in this chapter.

There are, however, two important caveats with this assumption. The first is that existing or planned systems did not anticipate supporting a large-scale personnel testing function and therefore may not have capability to do so. For example, the annual testing load may exceed projected DTF availabilities. The second is that there may exist functional gaps in the capabilities of the existing or projected systems that require the creation or acquisition of new capabilities.

Planning and integration are fundamental requirements, but it is essential that they be incorporated into a formal oversight and management plan. The first step is awareness but later, a more formal process, to include designation of overall Army organizational responsibility, must be adopted.

Dependability and Reliability

Testing for purposes of promotion places a premium on the requirements that the test delivery, scoring, and reporting system be accurate, dependable, reliable, and secure. Soldiers' confidence in the system is paramount for the success of the system. Simply put, operational "bugs" anywhere in the system are unacceptable. This means that system components must undergo rigorous trials and testing before implementation and that stringent standards be imposed and enforced. To do less threatens the success of the entire testing program.

Technology Updates

Rapid technology changes are axiomatic. It is also true that large-scale systems, such as those supporting Army operations, cannot retool for every change. Yet there has to be a middle ground between the possible and the practical. In a system that is developing, and has some period to go before committing to an operational incarnation, it is essential that technology changes be tracked and evaluated on a constant and consistent basis.

The remainder of this chapter documents the current knowledge and issues in the main three areas of technology application:

- Test development and delivery software
- Test delivery
- Administrative support

These areas were initially examined as part of Phase I of the project and documented in the report of that Phase (Knapp & R. C. Campbell, 2004 – see Appendix E). For the most part, the content and discussion in that report remains viable and is not repeated here. This chapter contains updates related to these areas as affected by subsequent developments and especially as impacted by the pilot tests conducted in Phase II. It must be kept in mind that, not only is technology very dynamic, but so too is the situation and posture within the Army, particularly given current operational commitments. What is factual or understood at the time of this writing may not be the case upon the reading. It is therefore important to follow-up this snapshot of the technology landscape and Army applications with continued updating through the appropriate resources.

Test Development and Delivery Software

In a revived Army testing system, the key to efficiency and success of the system will be software. The goal is for software that can do it all – author, administer, deliver and report on computerized assessments. Moreover, it must interface with the Army's automated personnel system and also, likely, with the Army's Distributed Learning System (DLS). This section outlines the specific capabilities requirements.

Availability and Technical Support

The Army currently uses some authoring systems, primarily in support of its distributed learning training function. However, since the Army does not currently have an inherent, large-scale test function, it is contended that test-specific authoring system software is a requirement. Further, the primary approach should be to take advantage of COTS rather than custom built software. However, the Army's test development and delivery system will be both large and complex. Therefore, even very capable COTS software will undoubtedly require some eventual modifications or customizations. Essential also will be the backing of an experienced and capable software support staff that can deliver both remote and on-site technical support anytime it is needed.

Maintenance Costs

The Army needs to acquire a long-term system capable of sustaining test development and administration over many years. Switching to other systems midway is not considered an option. While acquisition cost is a consideration, maintenance costs in an Army-wide system is an even larger issue. *Software maintenance* is defined as the totality of activities required to provide cost effective support to a software system to sustain the software product throughout its life cycle. It includes the modification of a software product after delivery to correct faults, to improve performance, or to adapt the product to a modified environment. Maintenance also incorporates training provided to users. Aaby (2001) estimated that 20% of maintenance costs are incurred in corrective actions and 80% of the lifecycle costs are due to system evolution. He defined four areas of software maintenance:

- Corrective – reactive modification to correct discovered problems
- Adaptive – modification to keep programs usable in a changed or changing environment
- Perfective – actions to improve performance or maintainability
- Preventive – modification to detect and correct latent faults

Ease of Use

There are two potential models for Army test development and delivery. The initial model (which focuses on Army-wide core tests) is centralized, both in test development and distribution. If MOS testing is eventually adopted, test development would be decentralized (to possibly 20 or more locations) although distribution would likely remain centralized. Both models are likely to exist simultaneously. Test developers (civilian test specialists, military SMEs, contractors) would be required to develop, modify, and maintain tests directly on the system. Analysts must be able to mine the system freely to extract results and performance data. Finally, the system must support interaction with the test taker and with test administrators.

Functions and Capabilities

The requirement is for a linked system that provides four testing functions:

- *Authoring* – must be able to support creation, modification, developmental tracing, and item banking of a variety of standard multiple-choice and non-standard (free response, matching, drag-and-drop, interactive) test items. Authoring must be Windows based and capable of supporting a variety of other software development tools (e.g., Word, Excel) and delivery means (e.g., Flash, Photoshop, JavaScript).
- *Delivery* – the primary anticipated delivery mode for Army testing will be via the Internet. There is a requirement to host potentially complex test content. The system must also support delivery via intranet or limited network servers. All delivery must be compatible with the anticipated portals (DLS suites) and must be secure.
- *Administration* – the system must facilitate scheduling, grouping, sorting, and classifying participants by a variety of criteria. Delivery of test items must be variable based on the sorting or identification parameters. The system must be capable of handling this administration for several hundred thousand participants annually.
- *Reporting* – results of individual tests must be reported by a variety of formats, both individually and aggregated. They must be extractable in different formats to facilitate various analyses and recording functions. Reporting must be compatible with various database tools and analytic software.

The specifics of these functions as they apply to a projected model of the Army test system will be discussed in detail later in this chapter.

Sharable Content Object Reference Model (SCORM)

SCORM is an initiative started by the U.S. Government that is intended to produce learning course content that is reusable. It consists of technical standards that enable web-based learning systems to find, import, share, reuse and export learning content in a standardized way. With SCORM, several parts, from different sources, can be put together by a server system (learning management system [LMS]) at the time of delivery. Most web content that has multiple facets uses hyperlinks to go from one content area to the other. In LMS based delivery, the LMS

is “smart” and directs the participant where to go automatically or limits where they can go based on what they have done before. SCORM is what makes sure that this feature works by providing standards for how a tool or application assembles or aggregates content, how it labels and packages that content, how the tool sends content to a central LMS, how it conveys that content to the central LMS, how the content should be ordered or sequenced, and how that content should be navigated through. SCORM is written primarily for vendors and toolmakers who build LMS and authoring tools so they know what they need to do to their products to conform with SCORM technically.

The U.S. Department of Defense (DoD) was assigned the lead for the U.S. Government’s SCORM program in 1999. The agent for the U.S. Army’s SCORM requirements is the Army Training Support Center (ATSC). For the past several years, SCORM has been an emerging and evolving package. *SCORM 2004* is the latest and most complete version of SCORM and the SCORM documentation is now considered “stable,” with future updates to be limited to lessons learned and operational adjustments. SCORM standards are applied within almost all U.S. Government agencies and LMS and authoring systems produced in response to U.S. Government procurement initiatives must be “SCORM conformant.” The same standards are widely applied in industry and in academia as well, and most vendors and producers are aware of SCORM conformity requirements and its associated documentation and tests. It is assumed that software systems used in support of an Army test system will conform to applicable SCORM requirements. DoD’s Advanced Distributed Learning (ADL) certifies SCORM conformance at their ADL Certification Testing Centers. A list of ADL Certified Products can be found at <http://www.adlnet.org/index.cfm?fuseaction=scormprod>.

Other Standards

Although as the U.S. Government’s primary program, SCORM has the highest interest, there are at least three other widely applied standards that are applicable and that reviewers should be aware of. These are:

- *AICC* – this is the Aviation Industry CBT (Computer Based Training) Committee. The AICC develops guidelines for the aviation industry in the development, delivery, and evaluation of CBT and related training technologies. The standards consist of technical recommendations called *AICC Guidelines and Recommendations (AGRs)*. There are nine AGRs, each addressing a different technology area. Training products can be designed to, or tested to, standard in one of more of the nine AGRs.
- *IMS QTI for CAA* – this is the Integrated Management System Question and Test Interoperability for Computer Aided Assessment Evaluation. This standard enables the reuse of question-based examination and assessment content between different examination development systems that comply. It facilitates examination authoring flexibility and portability by providing interoperability between different examination/assessment technologies.⁵

⁵ IMS QTI was a prime consideration for the U.S. Navy in their evaluation of technology-based testing. The Navy was looking for IMS QTI compatibility with their existing in-house Exam Development System (EDS) test development software.

- *IEEE LTSC* – this is the Learning Technology Standards Committee of the Institute of Electrical and Electronics Engineers. This standard is considered to be in development and intended as international standards by ISO/IEC JTC1/SC36. (<http://jtc1sc36.org/> and <http://ltsc.ieee.org/index.html>)

Although most standards address different issues and meet different requirements, there is also a certain amount of interoperability between most standards. For example, SCORM includes the following specifications in relation to these above standards:

- IMS Metadata, Content packaging
- IEEE Metadata Dictionary
- AICC Launch and Communication API

PerformM21 Software: History and Experience

In 2001, HumRRO initiated an in-house effort to identify a test development software authoring tool primarily to support projects that required military test development. Knowledgeable HumRRO staff conducted an Internet search of available software packages and eventually reduced the field to four promising candidates: FastTEST Pro®, Blackboard®, SMT-PCTest/SMT-Bank®, and Perception™ by Questionmark™. Staff then conducted a detailed evaluation of the proffered characteristics and capabilities of each of these packages. All of the packages presented advantages and disadvantages. In the end, Questionmark's™ Perception™ was selected. Perception™ offered the most acceptable combination of robustness, customizability, and affordability. Additionally, Perception™ had the advantage of being a more mature package, having had almost 10 years of developmental experience compared with only a few years development for the other candidates. HumRRO subsequently entered into a licensing agreement with Questionmark™ and had some minimal experience with the software when the PerformM21 project was initiated.

Concomitantly, the U.S. Navy's Naval Education and Training Professional Development and Technology Center in Pensacola, Florida was exploring the adaptation of electronic testing for enlisted promotion assessments. The Navy has an almost 50-year history of enlisted promotion testing and uses an in-house Navy developed software called Exam Development Software (EDS) for test development. EDS is used directly by Navy military SMEs in test construction. However, all test delivery is paper/pencil based (Knapp & R. C. Campbell, 2004). In 2002, the Navy began a study to assess the potential for computer-based delivery. They were also looking for a system that would interface with their existing EDS software. The Navy initially selected Questionmark™ Perception™ as well, based primarily on Perception's™ widespread commercial application and the Navy's existing use of Perception™ in end-of-course testing in Navy training programs (Baisden & Rymsza, 2002).

The Navy continued its exploration of computer testing with a more systematic review conducted in 2003. At that time, they examined in detail two leading software candidates: the previously selected Questionmark's™ Perception™ and OutStart Evolution®, a learning content management system (LCMS) that was used by the Naval Education and Training Command (NETC) as the primary content development tool for use with Navy e-Learning. The user

presentations offered in Perception™ were deemed to be desirable. Moreover, Perception™ conformed to the Integrated Management System Question and Test Interoperability (IMS QTI) Specifications for importing eXtensible Markup Language (XML) and eXtensible Stylesheet Language (XSL). Both XML and XSL are used in EDS, thus the Navy felt they had compatibility with their existing test development software system. In the end, the Navy concluded that there were not any benefits in moving to OnStart Evolution® over the previously chosen Perception™ (Shultz, Sapp, & Willers, 2004).

During Phase I and Phase II we used Perception™ extensively to author, administer, and deliver tests. The software package used consisted of three components: Question Manager, Assessment Manager, and Enterprise/Windows Reporter. We also used Perception™ to host the Internet delivery of the Army-wide tests in the pilot to about 600 Soldiers as described in Chapter 3. Overall, we found much to like about Perception™. It is a powerful and robust system with a lot of features including security, integration capability, and is generally backed by a knowledgeable and helpful support staff. Perception™ can integrate items programmed in Java or Flash as well as interactive graphics, video, and audio clips. This capability was explored, although perhaps not fully exploited, in the development of the so-called non-traditional items (free response, drag and drop, matching) used for the Army-wide core test. Moreover, the Perception™ test presentations were generally well received by the pilot test Soldiers – the presentations were compelling and the displays generally easy to follow and manipulate.

As a test item development and management tool, however, Perception™ does not currently provide features that would be required to support an operational test program such as that being envisioned for the Army. Necessary item banking features would include, for example, tools for identifying items based on the content area they cover and for storing historical information and data on each item. We also found it difficult to capture raw item-level data from Perception™ so that we could score non-traditional items in the ways we felt would be most psychometrically useful. The complex scoring strategies we required were not supported by Perception™ scoring protocols.

During Phase II (MOS test development) we sought to expand the boundaries of assessment by exploring simulations, off-line assessments, hands-on tests, product evaluations, and other assessment venues as described in Chapter 5. The further the departure from multiple-choice testing (including its non-traditional manifestations), the less amenable to development, administration, scoring, and accountability is a single test system such as Perception™ (or any other COTS). For example, the Patriot Air Defense Control Operator/Maintainer (14E) technical test described in Chapter 5 required a very large Java application that could not be created or hosted through the Perception™ system. Such limitations may simply be a fact of life, but the ultimate goal is to integrate, as much as is feasible, all test approaches and administration within the same system.

It is not within the requirements of the research to make input into eventual Army decisions on software to support an operational test system; such decisions are made following specified acquisition and procurement procedures to include meeting evaluative criteria. However, the experience of PerformM21 still should have weight and consideration in an eventual operational system. Perception™ is a good system for supporting certain test authoring

and delivery requirements. But the current available software was not designed to support the item authoring, scoring, and banking needs that will surface with the operational program envisioned for the Army. Since we originally selected the software to use in this feasibility research program, the available software options have evolved. Moreover, an operational system could make use of software that would have higher initial costs than could be incurred in a purely research effort, but that would better serve the needs of an operational system. We have initiated an updated review of testing COTS in Phase III to identify software alternatives with item authoring, item banking, and data capture features which might best serve the Army's operational testing needs.

Test Delivery

The goal of an Army test system is to have all Soldier testing functions – notification, preparation, test taking, and reporting/feedback – not only be computer-based but also conducted over the Internet. In this section, several issues pertaining to test delivery are presented.

Existing Army Capabilities – The Digital Training Facilities (DTF)

The Army is in the process of building a training network under its Distributed Learning System (DLS) initiative. The purpose is to reduce training costs while enhancing the delivery of training to the field, initially by moving schoolhouse instruction to a distributed mode, and later by enlarging the spectrum of what is offered and improving how training is delivered. There are four increments to DLS:

- *Increment 1: Digital Training Facilities (DTFs)* – these provide the infrastructure to deliver the courseware.
- *Increment 2: Enterprise Management Center (EMC)* – networks the DTFs and provides for performance management of the DLS.
- *Increment 3: Learning Management System (LMS)* – automates student management, content distribution, and learning content management functions.
- *Increment 4: Digital Deployable Training Campuses (DDTCs)* – provides for student surge requirements, support to areas of low Soldier density, or to remote locations such as during deployments.

Integral to the DLS is the fielding of the DTFs. DTFs are complete training suites consisting of computers, video tele-training (VTT) with two-way audio and video capability, network printers, local server, and Internet access. Each DTF suite includes 16 computer terminals with CD-ROM optical drive, Intel Pentium 2 or higher processor running Windows 2000 (or upgrade), monitor, keyboard, mouse, and headphones. DTFs support the Active Army and the U.S. Army Reserve (USAR). (A sister system, described later, supports the Army National Guard.) DTFs are run by local managers, usually operating under contract.

When fully deployed, there will be 274 permanent DTFs (plus 7 mobile DTFs) in 152 different locations. OCONUS permanent DTF locations are Alaska, Hawaii, Korea, Japan, Germany, Belgium and Italy. The 274 DTFs provide for 4384 “seats” at individual terminals.

These can support 197,500 students annually which is projected to increase to a capacity of over 2 million students when the LMS is fully implemented (PMO DLS, 2004).

Increment 3 – the LMS – has progressed at a slower pace. The Army has been working on developing a system that would provide for Army-wide automated training information management to include integration of numerous existing Army training systems and initiatives. This program is the Army Training Information Architecture-Migrated (ATIA-M) and would be incrementally developed over several years. In late 2003, the Army contracted with the software firm Saba to implement a DLS LMS by licensing Saba Learning, Saba Content, Saba Publisher, and Saba Virtual Learning Environment Extension. LMS fielding should start in 2nd Quarter FY05.⁶ The first year's fielding is projected to cover 18 DLS locations, primarily TRADOC school locations (TRADOC LMS Fielding Schedule, 2004).

DLS operates through the Army Training Requirements and Resources System (ATRRS). Essentially, ATRRS is an on-line training management information system (MIS) that is used to support the planning, programming, budgeting, and program execution of the training process. ATRRS is a function of the Army G1 (Military Personnel Management – Training Requirements Division) in support of its mission to integrate training management. ATRRS provides the following functions and information (HQDA Army G-1, 2004):

- *Course Information* – ATRRS maintains a database of course information on all courses taught for Army personnel. Information includes course administrative data, scope, and prerequisites.
- *Class Schedules* – specific yearly schedules for each training course.
- *Quota Management* – users can identify how many quotas they have to fill for each class, and fill date deadlines. Manages distribution of unfilled seats to other users. Allows trading of seats on-line between users.
- *Automated Training Application System* – application for training is entered and processed to make by-name and SSN reservations for training seats.
- *Student Information* – student personal and class status information is maintained on an individual file kept for historical purposes and to develop course utilization and attrition statistics.
- *Training Statistic Data* – during each phase of training development and execution, data is collected in an on-line database. The statistical database spans a 10-year time frame. Data can be categorized, summarized, compared and analyzed at a course level, component level, and on up to Army level of details.

The DLS LMS will be set up to interoperate ATRRS with AKO.

⁶ The existing DLS LMS is the TRADOC Educational Design System-Redesign (TREDs-R) which incorporates a COTS called LXR Test. Saba will replace this system. It is assumed the DLS will operate both systems for some period of time.

Army National Guard System – Distributive Training Technology Project (DTTP)

As indicated, the Army National Guard (ARNG) uses a system that is separate from the DTF framework described above. The ARNG system is the Distributive Training Technology Project (DTTP) that was started in 1995 as a public-private partnership to provide communications and training technology to both Soldiers and the communities that they serve. The National Guard Bureau (NGB), working in coordination with the individual state Governorships and/or state Adjutant General's Office, provides DTTP classrooms in armories, local schools, libraries, assisted-housing, and other community centers. The purpose is to provide a "technology hub" for individual communities that can be used as telecommuting centers, adult education centers, training centers for businesses and trade associations, and audio and video conferencing sites for tele-meetings. Access is provided on a fee-for-services basis. Individual classroom site capabilities include:

- Audio and video conferencing (live two-way)
- Computer-based instruction access
- Video programming
- Electronic mail and network access (Internet based)

There are currently about 315 individual multimedia classrooms and each classroom is based on a scalable design that can contain from 3 to 18 student workstations. There are classrooms in all 50 states, Washington DC, Guam, Puerto Rico, and the U.S. Virgin Islands. The DTTP classrooms are connected by a nationwide Asynchronous Transfer Method (ATM) network that can transfer voice, video, and data simultaneously. An updated listing of all site locations and status by state is available at [http://www.dttp.ngb.army.mil/About%20DTTP/Contact%20Information/State%20DL%20Classrooms%20and%20POC%20List%20\(HTML%20Version\).asp](http://www.dttp.ngb.army.mil/About%20DTTP/Contact%20Information/State%20DL%20Classrooms%20and%20POC%20List%20(HTML%20Version).asp).

The NGB is currently in the process of upgrading the DTTP system. Windows 95 and Windows NT 4.0 workstations that were part of the original deployment have largely been replaced. A new management system – the Guard Collaborative Learning Source (GCLS) – is being introduced to replace both existing management systems – the Integrated Information System (ISS) and the Site Administration Support System (SASS). The GCLS is how the DTTP deploys and administers training for both students and administrators. In many aspects, GCLS functions similarly to ATRRS (described previously) by providing a Distributed Learning Courseware Catalog for course dates, course prerequisites, scope, administrative data, course contact information, and location and enrollment criteria.

The stated goal of the Army is to provide 95% of the Army population with access to a distributed learning facility – DTF or DTTP – fixed location or mobile unit – within a 50-mile radius of their assigned location, by the year 2010. To meet this goal, the General Accountability Office (GAO, 2003) estimates they will need about 850 DTF/DTTP suites/classrooms. There are about 590 facilities in place or projected. All estimates are based on projections prior to the current Iraq deployments.

Test Delivery Issues and Goals

It has been assumed that an Army test delivery system would be (a) Internet delivered and (b) utilize (as much as feasible) existing Army systems. This has led to a working assumption that the DLS would provide the framework and the DTF/DTTP would be the portal for access to the testing. The pilot tests were conducted primarily at DTFs (see Chapter 3) and project staff members have been meeting periodically with DLS representation. It must be stressed that these assumptions and subsequent contacts have been unofficial; until various programs and systems have been tasked with supporting an Army test program the extent of compatibility, convergence, alteration, and expansion is only conjecture. Nonetheless, we have identified areas that bear attention as this aspect unfolds.

Bandwidth

Bandwidth is the measure – usually expressed in bits per second – of the rate at which information moves from one electronic device to another. It has become a central issue in the information age. By the late 1990s, copper wire-based bandwidth had been exhausted and the industry turned to revamping the infrastructure to respond to increased demands, essentially replacing an infrastructure that had been building for almost 100 years. Technology, primarily in the form of fiber optics, offers the most ready solution but the “rewiring” of the grid will not happen quickly or cheaply.⁷ Table 19 shows a comparative listing of current bandwidth linking capabilities.

But even as technology addresses some problems, the demand for bandwidth grows, especially in the training delivery field with its high dependence on interactive graphics and increasingly sophisticated simulations. In the Army-wide tests piloted for PerformM21, the bandwidth demands were modest. However, the areas explored in the MOS specific tests (see Chapter 5) only hint at the potential demands that will emerge in the future. It is a dictum of computer development that about every 18 to 24 months there is a doubling of the processing speed and emergence of some new capability. These new capabilities (and requirements) will undoubtedly affect the training and testing realms.

The DLS community is well aware of the limitations. GAO (2003) found a major insufficiency in bandwidth to support the services’ multimedia courseware for distributed learning, as well as a lack of sufficient funding to address the problem long term. As indicated in Table 19, absolute bandwidth capability is not the immediate problem; rather the cost and availability of the higher transmittal connectors, particularly in the military network system, given the Army’s geographical disposition, overseas locations, and requirement to support troops during deployments. To address the immediate problem, Army DLS uses Cisco content distribution networks to store as much courseware as possible at the local DTF site. Compact discs, local area networks (LAN) and regional servers are other options. Long-term, there is the pending availability of another Internet. But for the short-term future, bandwidth will continue to be a testing concern.

⁷ A T1 line in the U.S. typically costs about \$1,500 a month depending on location. Leasing Integrated Services Digital Network (ISDN) lines to conduct the Basic Noncommissioned Officer Course (BNCOC) distance learning in Korea is forecast at \$100 per hour or \$12,100 per class (Walcott, 2004).

Table 19. Current Bandwidth Capability Comparisons

Type Line	Capacity (bits/second)	Equivalency
Digital Subscriber Line (DSL)	64 kilobits	
Integrated Subscriber Digital Network (ISDN)	128 kilobits	2 DSL lines
T1 Line	1.544 megabits	24 DSL lines
T3 Line	43.232 megabits	28 T1 lines
Optical Cable 3 (OC3)	155 megabits	84 T1 lines
Optical Cable 12 (OC12)	622 megabits	4 OC3 lines
Optical Cable 48 (OC48)	2.5 gigabits	4 OC12 lines
Optical Cable 192 (OC192)	9.6 gigabits	4 OC48 lines

Note: From Fruedenrich (2004).

Capacity

The extent of Army testing is still an unknown factor. The original premise was to conduct annual testing of all Soldiers in grades E4 through E7. Table 20 shows the population estimates in those grades. To carry out a theoretical projection, if there was an assumption of both Army-wide testing and MOS-specific testing for all Soldiers, the requirement (based on a 3-hour per test facility access per Soldier) would approach 3.5 million hours per year. Of course, the reality, even in a fully operational test, is likely much more modest. Annual testing for everyone is likely not to be a recommended requirement. And tests may be more useful if targeted towards more specific career points (e.g., when E4 Soldiers reach promotion eligibility criteria). And finally, it is not a given that all MOS will conduct tests.

Nonetheless, the testing population is bound to be a formidable load. For example, the semi-centralized promotion system (to pay grades E5 and E6), which would likely be an initial focus of promotion test implementation, has, at any given time, about 177,000 Soldiers being processed in the system.⁸ At one time (and quite recently), the DTFs and the DTTPs were an underutilized resource, with very low usage rates.⁹ However, this has substantially changed in the last year. More and more distributed learning courses are coming online and deployments have caused substantial shifts to use of DTFs and DTTPs for a variety of training. Restructuring of the Noncommissioned Officer Education System (NCOES) requirements will make the DTFs the primary training source for common core training. MOS and pre-deployment training is also being conducted via DLS. It is unlikely that these are just temporary surges; if successful, they will undoubtedly usher in the forecasted wholesale shift to distributed learning. Under this scenario, the DTFs/DTTPs cannot, in their present capacity, be expected to meet the demands of an even moderately large testing requirement.

⁸ About 81,000 are Active Component and 96,000 are Reserve Component. The combined figure is for illustration only; semi-centralized promotions are processed differently in the Active and the Reserve Components and are not part of a single promotion system.

⁹ DTFs at one time reported use about 12% of available time. DTTP usage is more problematic – not all states report their usage and some of what is reported is questionable. Nonetheless, DTTP reports a 72% increase in total usage hours and 58% increase in number of users between October 2003 and October 2004 (DTTP, 2004).

Table 20. Population Estimates by Paygrade – Active and Reserve Components

Pay Grade	Active Army	Reserve Components	Total Force
E7	37,300	43,800	81,100
E6	57,500	67,500	125,000
E5	72,900	85,600	158,500
E4	103,000	121,000	224,000
Totals	270,700	317,900	588,600

Note: Adapted from Office of the Under Secretary of Defense, Personnel and Readiness (2003).

Security

Because of the high stakes nature of Army testing, security is a concern at all stages of testing – development and maintenance, delivery, and test administration. Additionally, personal and promotion information must be recorded. Since all will be computer-systems based and much online, security concerns need to be addressed system-wide, not just at the point of administration. The manual system of test accountability involves distribution through designated test control officers (TCO), numbered accountability of test instruments, and an auditable record system of test transmittal, storage, and possession. A similar system must be duplicated in an electronic test system.

The current DTF suite layout and configuration is conducive to a secure test administration. The suites are easily monitored by a single proctor and examinees cannot easily view other's screens. Computer installation is such that test takers cannot readily access the backs of the computer processing units (CPU) and none of the front plates or peripherals had universal serial bus (USB) ports.¹⁰ The delivery software used – Questionmark's™ Perception™ - offers a utility called Questionmark Secure which can prevent users from printing questions, using the right-click on the mouse, saving the HTML, viewing the source, opening new applications, task switching, and freely exiting the application.

One of the security-related concerns has to do with currently existing firewalls and protections erected in Army computer systems. GAO (2003), in its review of the Army distributed learning program, found "emphasis on securing networks may impede learner's ability to access education and training anytime, anywhere" (p. 30). In our own pilot tests, we experienced some access problems that are attributable to local security precautions and protections (see Chapter 3). Moreover, local conditions may affect test development requirements. During development of the MOS tests, we attempted to host remote test item reviews using an Internet linkage service but firewalls at one major Army installation would not allow this access.

Proctoring

Inherent in the design for Army testing is a requirement for test proctors – certified individuals who would be present for each test administration to verify examinee identification

¹⁰ Our pilot and observational experience with the ARNG DTTP facilities was much more limited. Since these configurations tend to be less standardized, no overall conclusions about DTTP test-taking security are included.

and monitor the test taking activities. In effect, proctors would be de facto test administrators and must work closely with site technical personnel during the conduct of the test. It is not anticipated that DTF or DTTP site administrators would be test proctors.

Portals and Hosting

During the Phase II pilot testing described in Chapter 3, Soldiers accessed the test over the Internet, primarily through DTF portals, to a host site provided by Perception™. During the pilot tests, at nine different geographical locations, technical problems were pervasive and, in some cases, severe. Problems (see Chapter 3) were classified as:

- *Login problems* – Soldiers had to be registered in the LMS in order to use the DLS facility. This was not a faultless process. Small errors in Soldier spellings or incorrect email address could stymie the login process. The process was very uneven across sites.
- *Computer specific problems* – problems ranged from insufficient software to inability to load simple graphics to insufficient memory. Again, there were many variations across the sites.
- *System-wide problems* – multiple machines would repeatedly freeze, crash, or time out. Some users would be inexplicably bumped off the system. Systems would not accept submission of data. Sometimes it was impossible to tell what had been recorded and what had not.

Sometimes the source of the problem could be specifically identified; sometimes there was disagreement over the root cause. In any case, hosting was not perfect, with multiple instances of examinees being dropped off-line during test administration and other times in which the hosting system was down for maintenance. In an operational Army-wide, worldwide, 24-hour access system, such occurrences would be intolerable. This may cause reconsideration of the “one-source, one-system” approach to choosing testing COTS. Increasingly, it is a practice in the commercial testing community to use specialized software for specific functions (e.g., authoring, item banking, analysis) and to create customized software for test administration, reporting, and linkages between various components. It may evolve that the choice for the Army is to develop a hybrid mix of COTS, internal, and custom built components and software. Additionally, development needs to include investigation of alternatives to the aforementioned delivery and hosting means to include server-based, and even CD-based administration. While the emphasis should be on making the proposed system work properly and effectively, effort needs to begin on exploring alternate systems both as primary and as back-up capability.

A further concern is when the test model expands to include different types of testing such as high fidelity simulations or hands-on tests as explored in the MOS test approaches in Chapter 5. This certainly would involve hosting by alternate means and also testing at locations other than DTFs/DTTPs. A test program that expands beyond these features and facilities is certainly possible, but it does raise issues of recording, reporting, and integration of multiple test venues. These challenges will be further explored in Phase III when we pilot test some of the MOS tests.

Administrative Support

There are many components to a successful test system, some that exist only to support the testing and others that must interact to support the system. From a technical requirements standpoint, it is essential to keep in mind the interactive nature of many of these supporting functions. This section will attempt to outline that overall view. This is depicted in Figure 7. Salient features of that depiction are discussed in this section.

Defense Integrated Military Human Resources Systems (DIMHRS)

In the mid-1990s (following the Gulf War), DoD became aware of a serious disparity and many problems caused by the different services, and in particular by the Reserve Components, having separate and unique personnel and pay systems. Subsequently, and after much study, it was decided to establish a single military human resources system with a single logical database for all components, to allow integrated personnel and pay processes at all echelons during both peacetime and war. The Joint Requirements and Integration Office was established to coordinate the effort and to evaluate and select a COTS product for the system. The COTS product selected was PeopleSoft®.

Implementation of DIMHRS for the Army is to be a migratory process, using a progressive approach to replace legacy systems. For the Army, there are eight primary systems to be replaced including Standard Installation/Division Personnel System-3 (SIDPERS-3), Electronic Military Personnel Office (eMILPO), and the Total Army Personnel Data Base (TAPDB).¹¹ Individual legacy systems will be maintained until all of the legacy systems are replaced (Department of Defense, 2002).

Soldier Eligibility and Test Notification

Army promotion is one of the functions that will be processed and administered under DIMHRS. Substantial change will probably be forthcoming in the administration of the Army promotion function. Nonetheless, it is essential that any Army test system be integrated into that function and into DIMHRS. The requirement is that, based on criteria yet to be identified, the DIMHRS will identify, on a predetermined basis, those Soldiers eligible or required to be tested in upcoming test periods. This identification must be followed by an automatic notification, both to the Soldier and the Soldier's commander, of the testing requirement. The notification must include a time and a testing location based on the system's identification of Soldier's geographical location. After testing, there must be feedback through the notification system that the testing was accomplished; if not, the notification system must reschedule and re-notify the Soldier and the commander for make-up testing.¹²

¹¹ The Army maintains different SIDPERS, eMILPO, and TAPDB systems for the Active Component, the ARNG, and the USAR.

¹² This is assuming an Army-wide test. If an MOS had a job specific test, it is anticipated this will be handled as separate test but with similar notification procedure. Thus, some Soldiers would have two separate test requirements.

Score Reporting

After the Soldier is tested there will be a delay period while the test items are verified and individual test scores computed. At least for Soldiers in the semi-centralized promotion system (to pay grades E5 and E6), test scores will be converted to promotion points. For example, if the promotion test were worth 200 promotion points, a Soldier would receive a scaled number of points from 0 to 200 based on their test performance. However, since promotion is by MOS and pay grade (and by component), it is also anticipated that Soldiers will want to know their standing relative to the Soldiers with whom they are competing. Therefore, it is a likely requirement that almost all of the population in a group (by pay grade, MOS, and component) will have to be tested before results can be provided.¹³

It is anticipated that Soldier reporting will be fairly straightforward. Soldiers will need to know both points standing and some relative performance information. They will also need to get some diagnostic feedback of performance in areas of the test (such as by Common Tasks, Leadership, Training). Feedback should be automatic and transmitted or posted electronically. Individual scores and points will have to be transmitted back to DIMHRS for posting to individual records and for integration with other promotion criteria and the production of promotion standings and promotion lists.

Feedback to units is potentially more complicated. Units will probably need individual results but they will primarily be interested in roll-ups by unit and by category (grade and MOS) as an indicator of readiness. Since testing will likely be spread throughout the year, roll-ups will have to be continuous, likely on a quarterly basis, but perhaps more frequently and perhaps on an on-demand basis. Roll-ups should also reflect personnel losses and gains and detail in the roll-ups should include test content areas. It is anticipated that roll-ups would be at the battalion level but other aggregates may also be required.

Analysis and Management

The data collected during testing becomes the wellspring of information on which a variety of programs and functions depend. Some of these are depicted in Figure 7. Of immediate concern in any test cycle is the identification of invalid items and identifying indications of test compromise and cheating. Developers need results data to plan for test modifications and to guide them in subsequent test development efforts, including item analysis to identify how items on the test are interacting with each other. Additionally, some items on each test will likely be administered as trial or as items that are not scored for record. Websites and test guides that help Soldiers prepare for the tests and to conduct a self-assessment of their test readiness need to be updated based on testing performance. Training management and personnel management decision makers need results to guide subsequent policy decisions.

All of these needs require different information. Some is immediate and time sensitive; some is cumulative and historical. The system must be responsive to sorting, aggregating, identifying, and even projecting data under a variety of criteria and with a minimum of human intervention. All this must be done while protecting the integrity of the system and of the data.

¹³ Promotion points are not used in the centralized promotion system (to E7, E8, E9). However, comparative standings with others in like MOS, pay grade, and component, are equally important in centralized promotion decisions.

It must be made clear that the concept of how the test system will function is tentative and changeable. Much will depend on policy and implementation planning that can only be made once an implementation decision and structure is decided. Yet, system planning needs to advance. Administration support requirements surrounding the actual test develop and delivery software could be a significant factor in the success and utility of the system and must be planned at the earliest possible time.

Summary and Conclusions

Technology matters are the most significant single issue that an Army test system implementation will face – the success and cost-effectiveness of the whole test program is dependent on achievement of technology goals. The technology requirements are emerging and will continue to materialize, but planning is paramount and it is essential to target areas that require focus and attention. What exists now is a reasonable model, given current thinking and information. It is subject to change but provides a practical starting point. The primary summation issues are:

- Technology costs and effectiveness are difficult factors to determine but are essential to gauging the overall costs of an Army test program. Phase III of PerformM21 should help to identify the parameters of such an assessment but more system definition may be required to make reasonable projections.
- Questionmark's™ Perception™ has been the COTS of choice for test development and test delivery during the research phase. As the Army moves toward operational implementation of a testing program, however, it will be necessary to once again evaluate all available COTS software alternatives. Of particular concern is the need to identify software that will allow the item authoring, delivery, scoring and item bank management features and flexibility required to support an operational program of the type we have discussed here.
- Use of the DLS (DTF/DTTP) facilities as the portal for test administration is the only logical avenue to pursue at this point; establishment of a separate test system infrastructure is not considered a viable alternative. However, the existing facilities likely cannot support operational testing without expansion of capability. Furthermore, there may be aspects of MOS-specific testing that are not supportable from the DTF/DTTP framework. Informal coordination has progressed about as far as possible. The Army DLS community must be formally tasked with exploring the requirements of being assigned the test support mission.
- Reliability and dependability at the point of delivery to the Soldier must have a goal of system perfection – system failures that negatively impact the test experience cannot be tolerated. This 2-3 hour exposure is what the Soldier sees and remembers and where impressions will be formed. Soldiers must have faith in the test system for Army testing to succeed.

CHAPTER 5: DEVELOPMENT OF PROTOTYPE MOS ASSESSMENTS

Deirdre J. Knapp, Karen O. Moriarty, Roy C. Campbell, Chad Van Iddekinge, Lee Ann Wadsworth, Alicia Sawyer, Masayu Ramli, Andrea Sinclair, and Carrie Noble¹⁴

Introduction

It is necessary to assess both Army-wide core competency areas *and* MOS-specific technical competence to capture a full measure of a Soldier's competence. MOS-specific assessment presents a particular challenge for the Army for many reasons, including the following:

- Currently there are roughly 180 MOS, each with multiple skill levels.
- There is insufficient commonality across the technical requirements for MOS (even those within the same Career Management Fields) to permit using the same assessment for multiple MOS.
- MOS requirements and the MOS structure evolve over time, making them a moving target for assessment.
- For many MOS, there is considerable variation in what Soldiers do in their particular assignments (i.e., unit or assignment specific equipment or missions).
- There is an historical preference associated with hands-on testing that reinforces the interest in having highly realistic assessments.

As mentioned in Chapter 1, it is not particularly difficult to design and maintain a high quality assessment. But it is difficult to do so in a cost-effective fashion under the types of circumstances listed above.

Research Strategy and Resources

The purpose of the MOS assessment portion of the PerformM21 research is to tryout ideas for designing an MOS-specific assessment program that will meet the needs and interests of the Army and individual Soldiers. We selected five MOS and collected job analysis information for each. Although we originally planned to use the job analysis information to design and develop a fairly comprehensive prototype assessment package for two or three of those MOS, we instead developed prototype test items for all five MOS. This strategy allowed us to try more assessment methods but resulted in prototype assessments that do not comprehensively cover MOS requirements.

At the outset of the research, we targeted our work to the design and development of assessments that would be suitable for administration to Soldiers eligible for promotion to Sergeant (E5). We operationalized this as E4 Soldiers with approximately 3 years time in service. It is important to note, however, that lessons learned in this experience would be applicable to other pay grades as well.

¹⁴ Considerable support for the work described in this chapter was provided by Shelly West, Shonna Waters, Eduardo Jimenez, Tonia Heffner, and Pam O'Quinn.

Although we were not able to mirror exactly how job analysis and test development work would be accomplished in an operational program (e.g., we had less time and SME input than would be required for an operational assessment), we had a variety of resources to support our work. SME input was provided by small groups of NCOs tasked through ARI's research support request process. The basic request was for three visits to the Advanced Individual Training (AIT) schoolhouse for each MOS. For each visit, the schools were asked to provide 16 instructors and 12 students for 8 hours each across 3 days. Some proponent organizations were also able to provide additional support (e.g., the 31B proponent gave us permission to administer a web-based survey to a sample of Soldiers) and the ATPAT also provided assistance. Note that, once the five target MOS were identified, the ATPAT was expanded to ensure representation from each of these MOS.

As discussed later in this section and in prior related research (Rosenthal et al., 2005), a high quality multiple-choice examination is likely to be an important component of any MOS assessment package. ARI has sponsored prior research in which such tests have been developed for a variety of MOS. The Select21 project team recently developed such tests for six MOS, which are shown in Table 21 (Knapp, 2003). The Project A research team developed tests for 20 MOS that are somewhat more dated given they were last administered in 1991 (J. P. Campbell & Knapp, 2001). All of these tests help serve as prototypes for what a major component of an MOS competency assessment would look like and some of the items on these previously developed tests were used, updated, or otherwise modified for incorporation into the prototype tests described in this chapter.

Table 21. Select21 MOS

MOS	Name
11B	Infantryman
19D	Cavalry Scout
19K	Armor Crewman
31U (now 25U)	Signal Support Systems Specialist
74B	Information Systems Operator/Analyst
96B	Intelligence Analyst

Selection of Target MOS

We began this part of the research by working with the ATPAT to identify five MOS to examine. Determining which MOS would form our sample was a decision based on several considerations:

- Select MOS that are highly rated on overall performance criticality requirements.
- Select MOS expected to be relatively different from each other with regard to the measurement methods that would be most applicable for high quality assessment.
- Include MOS with current occupational analysis information.
- Include MOS for which there are related civilian credentialing programs.
- Include MOS that are in a state of transition.
- Include MOS that are represented by a proponent organization that is interested in supporting the PerformM21 project, particularly with regard to providing SME support.

We relied on the ATPAT to make sure the target MOS met the first requirement of relative mission criticality. With regard to the second, Rosenthal et al. (2005) identified seven clusters of MOS based on which assessment methods (e.g., fact-based multiple-choice tests, situational judgment tests, hands-on tests) would likely be best for measuring performance. We tried to sample from these clusters to ensure diversity in subsequent measure prototyping work. We used the remaining criteria as a checklist when we considered candidate MOS. Since most MOS do not have current occupational analysis information, this was not a major determinant of the final selections.

Table 22 lists the five target MOS that were selected. Both 14E and 19K were selected because they held promise for relatively high fidelity simulation testing. With 19K, we wanted to explore adapting the high fidelity simulators developed for training for use in assessment. The possibility of dual-use technology could make high fidelity testing potentially more affordable. The 31B MOS was selected in part because military police (MPs) have played a large role in peacekeeping activities, which is a major component of current deployment activities. This MOS also has the complication that there are two types of assignments in this MOS (garrison law enforcement and peacekeeping/combat support) that may require different competencies. Finally, both the 63B and the 68W MOS were undergoing consolidation and were ideal for trying principle or systems-based (as opposed to task-based) testing. They also both have related civilian credentials. These five MOS are quite different from each other and together provide an array of opportunities and challenges to tryout new methods for collecting job analysis information and methods for assessment.

Table 22. PerformM21 Target MOS

MOS	Proponent Location
14E Patriot Air Defense Control Operator/Maintainer	Fort Bliss, TX
19K Armor Crewman	Fort Knox, KY
31B Military Police	Fort Leonard Wood, MO
63B Wheeled Vehicle Mechanic	Aberdeen Proving Ground, MD
68W Health Care Specialist	Fort Sam Houston, TX

Table 23 shows the site visits and the SME support that was provided during the course of the job analysis and test development work. In addition to providing job analysis and test development input, SMEs also were briefed on the concept of competence testing and given the opportunity to comment. In addition to SME workshop participants, each MOS proponent was asked to provide a point of contact (POC) to facilitate communications and cooperation among research staff and participating Army personnel. In addition, the proponent POCs were invited to join the ATPAT.

Overview of Chapter

The next two sections of this chapter discuss job analysis and test design, and test development. These discussions focus on MOS testing in general, using our work with the five target MOS for illustration. These discussions also refer back to material in Chapter 3 because many issues are similar whether considering an Army-wide core examination or MOS-specific assessments. The chapter closes with a focused discussion of each of the five target MOS, which recaps our experiences with each and provides a vision of what a full-scale operational

assessment might entail. This summation illustrates the uniqueness of MOS testing and the requirement to take a diversified and individualized approach while still adhering to core testing principles and parameters.

Table 23. Summary of Site Visits and Subject Matter Expert Participation

MOS	Month	E-6	E-7	E-8	Other	Total
Patriot AD Operator/Maintainer (14E)	April	8	3	0	0	11
	July	3	3	0	1 officer	7
	<i>Total 14E</i>	<i>11</i>	<i>6</i>	<i>0</i>	<i>1</i>	<i>18</i>
Armor Crewman (19K)	February	1	18	0	0	19
	August	4	13	0	0	17
	November	5	7	0	0	12
	<i>Total 19K</i>	<i>10</i>	<i>38</i>	<i>0</i>	<i>0</i>	<i>48</i>
Military Police (31B)	April	9	15	0	0	24
	July	6	15	0	0	21
	October	8	15	0	0	23
	<i>Total 31B</i>	<i>23</i>	<i>45</i>	<i>0</i>	<i>0</i>	<i>68</i>
Wheeled Vehicle Mechanic (63B)	August	11	5	0	9 civilians	25
	November	5	1	0	0	6
	<i>Total 63B</i>	<i>16</i>	<i>6</i>	<i>0</i>	<i>9</i>	<i>31</i>
Health Care Specialist (68W)	March	13	14	0	1 unk	28
	June	12	8	1	1 unk	22
	Aug/Sep	9	14	0	0	25
(Internet/phone conference)	November		3			3
	<i>Total 68W</i>	<i>34</i>	<i>39</i>	<i>1</i>	<i>2</i>	<i>76</i>

Note. All activity took place in 2004. Unless otherwise specified, we met with SMEs at their proponent location. 63B site visits were at Fort Jackson, SC because that is where training was being conducted.

Job Analysis and Test Design

The type of job analysis information required to support test design and development is not the same as that required to support training programs, which is one reason the Army's current occupational analysis program will not satisfy the needs of a competency assessment program. Moreover, the type of job analysis information needed to support test development varies depending upon what measurement methods are going to be used. For example, development of a situational judgment test requires what is known as a critical incident analysis (Flanagan, 1954). Rosenthal et al. (2005) proposed conducting an inexpensive, highly standardized, preliminary job analysis to identify what measurement method(s) would be most suitable for an MOS assessment process. Subsequent, larger-scale job analysis procedures would be tailored to meet test development requirements.

Preliminary Job Analysis

We implemented the Rosenthal suggestion by using the first site visit (for all MOS but 63B) to collect data on four sets of generic job descriptors – that is, work characteristics that are not job-specific in nature. These were largely drawn from the taxonomies that are part of the Occupational Network (O*NET) database maintained by the U.S. Department of Labor

(Peterson, Mumford, Borman, & Fleischman, 1999). Following is a brief description of each set of descriptors.

- Generalized work activities (GWAs) are a set of statements that are task-like in nature, but which are more abstract than the detailed job tasks typically used in the Army. Examples are “inspecting equipment, structures, or materials” and “evaluating information to determine compliance with standards.”
- Cognitive complexity indicators, such as judgment/problem-solving, information intensity, and systems thinking.
- Work context descriptors, such as requirements for vigilance, attention to detail, social interaction, and performing under extreme conditions or time pressure.
- Declarative knowledge requirements – these are akin to “textbook” knowledge and include a number of generic skills (e.g., reading, writing, listening, interpersonal).

In addition to collecting ratings on these job descriptors, we also asked the SMEs to generate critical incidents. These are actual examples of particularly effective or ineffective performance that we used as a tool to help SMEs think about their MOS requirements. Such incidents also can form the basis for test content.

We conducted an initial trial of these analytic measures at a workshop with our first set of SMEs, who represented the 19K MOS. Since we were interested in development of a test suitable for administration to E4 Soldiers interested in promotion to E5, we asked these SMEs to describe the job requirements of an “E4 eligible for promotion.” They used a 5-point low to high scale to rate the GWAs and a 3-point low to high scale to rate the other job descriptors listed above. The SMEs complained that importance and frequency were confounded in the wording of the scales. They also had trouble distinguishing an “E4 eligible for promotion to E5” from an E5; subsequently we changed the protocol to refer to an “E4 with about 3 years time in service.” We also found that ratings for each descriptor were uniformly high. We modified the forms so that frequency and importance each had its own 5-point scale and asked subsequent SMEs to provide preliminary ratings followed by a second rating after a discussion period. We theorized that a discussion along with the scale modifications would introduce more variance in the ratings.

In addition to collecting ratings for each MOS on these descriptors, we developed a list of job tasks using available resources (e.g., MOS training manuals). During the initial site visits, we had participants review and revise these task lists. In some cases, they sorted the tasks into the GWAs as a strategy for quickly identifying missing or redundant tasks.

Appendix D (Tables D1 through D4) shows the post-discussion results of ratings collected on the modified forms for three MOS. The 19K ratings are not included because the rating process was changed based on the initial tryout with this group and ratings were not collected for the 63B group because the MOS was in a state of transition. However, for the other three MOS, the GWA ratings show some intuitively sensible results. For example, the GWAs that are mostly supervisory in nature (e.g., making decisions, monitoring resources, scheduling activities) tended to get uniformly low ratings as would be expected for nonsupervisory

personnel. Repairing and maintaining mechanical equipment is quite important for the 14E MOS and less so for 31B and 68W. In contrast, 14E Soldiers have less need to communicate with people outside the Army than Soldiers in the other two MOS.

Despite our efforts to encourage SMEs to be more discriminating, ratings for most of the job descriptors still tended to be relatively uniformly high, although there were some expected differences between MOS. For example, the 14E MOS has extremely high requirements for systems thinking whereas the ratings for the other two MOS are more moderate.

In retrospect, we believe there is some value in using these job descriptors to help SMEs think about job requirements in a broad fashion (and particularly for thinking of requirements not just in terms of detailed job tasks), but ratings alone will not lead to good decisions about measurement methods. Also, our focus on E4 Soldiers with 3 years time in service may have been unnecessarily restrictive. In an operational job analysis, it is likely that it would be preferable to use survey data collected from a large sample of E4 incumbents, dropping the data only of those E4s with very little experience (e.g., less than 1 year time in service).

Identification of Test Methods

In this research effort, and we expect in an operational setting, decisions about measurement methods were ultimately made by the proponent SMEs and POCs with guidance and assistance from testing professionals to help them understand the measurement choices possible and their ramifications. For purposes of the research, we approached this issue at two levels: (a) what test methods would be recommended for an operational assessment and (b) what test methods would we tryout for purposes of this research.

To aid in the decision-making process, we used matrices that crossed measurement methods by the five target MOS. We were not able to try every method (e.g., we found no opportunities for on-the-job assessment), and wanted to learn as much as we could from each prototype we developed. So, for example, the job knowledge test we developed for the 63B MOS uses a systems-based test blueprint to help demonstrate how this type of design might work better than the task-based blueprints traditionally used for Army testing. Table 24 shows a matrix that summarizes how the project team used the five MOS to tryout different assessment methods.

Table 24. Assessment Methods by MOS

Method	14E	19K	31B	63B	68W
On-the-job assessment (automated capture of performance during training or other job application).					
Expert evaluation of actual work products			X		
Hands-on work sample tests		X			
Computer-based simulations	X		X		
Multiple-choice "simulations"				X	X
Multiple-choice situational judgment test			X		X
Multiple-choice test (incorporating visual aids and audio/video clips and non-traditional item formats such as matching, ranking, and drag-and-drop)		X		X	

Note that in an operational assessment program, we would expect virtually every MOS to include a high quality multiple-choice test. As previously mentioned in this report, prototype assessments of this variety are already available. Most notably these include the PerformM21 Army-wide prototype test (see Chapter 3) and tests recently developed for the Army's Select21 project (Knapp, 2003). In our work, we are moving beyond what was previously done in several ways. The 63B test is designed around the principles of mechanics rather than vehicle-specific tasks. This is consistent with how AIT training for this MOS has shifted. For 19K, we took the test developed for the Select21 project and added or modified items to incorporate higher fidelity presentation of test content (e.g., more and better graphics).

We decided to develop situational judgment tests for two of the target MOS (31B and 68W). In other contexts, this has been a very useful method to address aspects of the job that require informed judgment in addition to textbook knowledge (McDaniel, Morgeson, Finnegan, Campion, & Braverman, 2001) and these two MOS seemed to have a particular requirement for such judgment that goes above and beyond more general Soldiering requirements. Recall from Chapter 3 that the Army-wide assessment also includes a situational judgment test (i.e., the LeadEx). Development of these tests requires a special step in the job analysis phase to obtain problem scenarios through generation of critical incidents.

SMEs in the 19K MOS were particularly interested in integrating a hands-on test component to their vision of an operational assessment. Indeed, it is likely that a high quality hands-on test would be a desirable element of the assessment package for most MOS (Rosenthal et al., 2005). As with fairly traditional multiple-choice tests, however, we already know how to create and score a good hands-on test. The problem is the practical limitations of this assessment method. Therefore, instead of developing prototype hands-on tests, we explored the idea of hands-on testing without actually developing any new tests.

The 14E MOS was selected in large part because it seemed a good candidate for designing a relatively high fidelity computer-based simulation that would rival the best possible hands-on assessment. The simulation we have developed is based on presentation of a situation involving operation of the Patriot missile system rather than seeking to test isolated tasks. We selected the 19K MOS in part because we thought it might be possible to adapt existing high fidelity tank simulators used for training to meet testing requirements. This was not possible for several reasons, including the lack of availability of training simulators for test purposes and the expense associated with programming for an assessment application. We were, however, able to use graphics and animation originally developed for training applications to design lower fidelity simulation-based assessments.

For some MOS, it might be possible to score actual work products as an element of an operational assessment. We explored this idea with the military police (31B), focusing on their incident report form.

Full-Scale Job Analysis

As part of the PerformM21 research, we did not conduct job analysis work to the scope that would be necessary to support an operational assessment program. Had resources (time, money, personnel) permitted, we would have conducted additional SME workshops and collected input from a much broader sampling of Soldiers using an automated job requirements

survey. Instead, we used the second and third site visits with proponent SMEs to collect job information to support development of the prototype measures. The procedures we used and data collected are discussed further in the next section on test development.

An Incumbent Survey

Although the survey was conducted after prototype assessments had been developed, we were able to develop and administer a task analysis survey to a sample of 31B Soldiers. As discussed in Knapp and R.C. Campbell (2004) and again in Chapter 2 of the present report, the Army's existing occupational analysis system (known as ODARS) does not currently provide the information required to develop effective assessments. It would be desirable, however, to leverage this existing system inasmuch as possible to collect the necessary data. We tried this by starting with the latest available task list for the 31B MOS and using ARI's AUTOGEN survey development and delivery system. ARI developed AUTOGEN for data collection in support of the ODARS program. Below we briefly describe the survey content, design, administration, and results.

Content. Our goal was to show how an occupational analysis survey could be used as a basis for determining what job content should be tested. It could identify the most important aspects of the job to consider covering with the resource-intensive assessment methods (e.g., simulations or hands-on tests). Survey results can also be the basis for developing a blueprint for a multiple-choice job knowledge test. As discussed in more detail in the next section of this chapter, a test blueprint shows the percentage of items on a test that should tap different aspects of the job domain. It is also helpful to develop a test blueprint so that it supports calculation of subscores on different parts of the test that can be used to provide diagnostic feedback to examinees.

Although we would have preferred to conduct a survey that focused on MOS-specific KSAs instead of (or in addition to) job tasks, we did not have the resources to develop a comprehensive list of KSAs specific to the MPs. To develop the survey, therefore, we began by reviewing the 140 Skill Level 1 and 2 tasks from the current 31B Soldier training manual publications (STPs). Because highly detailed task statements are not required for test development, we looked for ways to combine highly similar tasks into somewhat broader sets of tasks¹⁵. This also greatly reduced the number of tasks on the survey. For example, three original tasks "Load an M9 pistol," "Unload an M9 pistol," and "Engage targets with an M9 pistol" were combined into a single task statement ("Load, unload, and engage targets with an M9 pistol"). The development of the revised task list involved several iterations and internal reviews. We also sought input from 31B SMEs during the second data collection trip to Fort Leonard Wood. Specifically, we asked ANCOC instructors and students to review the task list and suggest tasks to add, delete, or revise. Lastly, we asked our 31B proponent POC to review the final task list and provide suggestions for improvement. The final list comprised 106 tasks.

Once the task list was finalized, we developed categories to help organize them. We began with the task categories from the 31B Soldier Training Publication (STP) and revised them based on the new task list. For example, the "Law Enforcement" category comprised many more tasks than did the other categories, and thus we divided this category into several, more

¹⁵ We understand this approach is not consistent with that prescribed by TRADOC Pam 350-70, but it is consistent with the goals of this testing (as opposed to training) oriented research effort.

homogeneous subcategories of tasks (e.g., "Traffic Control," "Apprehend Subjects"). The task categories underwent several internal reviews and then a final review by our proponent. The final list included 18 task categories.

Given our interest in supporting development of a viable test blueprint that can be used as a basis for test development and providing feedback to examinees, we wanted a smaller number of categories that would not vary greatly in size or overlap with each other, and would be meaningful for feedback. Therefore, we combined similar categories of tasks into a more manageable set of higher-order task categories. As with the tasks and lower-order task categories, the initial set of higher-order categories was developed internally and then reviewed by our proponent. The final set included five higher-order categories (e.g., "Respond and Apprehend"), each comprising 2-5 subcategories of tasks. The idea was to have survey respondents assign weights (summing to 100%) to task categories in a way that best reflects their relative importance to job performance.

Design. Once the content of the survey was finalized, we worked with ARI's Occupational Analysis Office (OAO) to design it. The content and aim of the survey differed from the typical OAO survey. That is, the primary goal of this survey was to evaluate the AUTOGEN system for collecting job information to develop a test blueprint. In contrast, AUTOGEN is typically used to collect occupational data to inform training and development activities (e.g., to determine which job tasks should receive the most attention during training). Thus, several customized changes were required to develop the current survey. For example, different instructions and rating scales were needed. In addition, the weighting of higher-order task categories, which took place at the end of the survey, had to be customized.

The final survey comprised three main sections. The first section included an introduction that described the purpose of the survey. Respondents were also asked to provide various pieces of background information in this section, including their rank, component, and the number of months they were deployed during the past year. Next, respondents were asked to make two judgments about each of the 106 tasks. Specifically, for each task they rated (a) its importance to performance of a 31B Soldier and (b) the percentage of 31Bs that perform the task. In the final section of the survey, respondents were asked to distribute 100 points across the five higher-order task categories in a manner that represents their relative importance to 31B performance.

Administration. The survey was administered using the OAO's existing survey procedures. Specifically, the Total Army Personnel Database was queried to obtain a list of survey respondents. The survey was then sent via electronic mail to all E4-E6 Soldiers within the 31B MOS ($N = 10,057$). To access the survey, respondents simply clicked on a link that took them to the introduction page of the survey, which was housed on the OAO's occupational survey server. The survey was available for Soldiers to complete for approximately one month.

Results. The survey administration period concluded during the writing of this report, and analyses will commence upon receipt of the survey database from the OAO. Results will be used for two main purposes. First, as discussed, to demonstrate how the data could be used to design a test blueprint for a job knowledge test. We will also provide the results to the proponent school to be used as they see fit.

Conclusions. We believe that an online process such as this could be valuable for collecting occupational information to develop competency assessments. For example, online technology allows researchers to collect data from large samples of individuals in a more timely and cost effective manner than face-to-face methods. Such processes also enable Soldiers to complete surveys at a time that is most convenient for them.

Despite these and other advantages of online data collection, the OAO currently is not setup to support such efforts. For example, the AUTOGEN software is not easily adaptable to survey formats that differ from the typical OAO survey. Further, it does not appear that the OAO is sufficiently staffed to take on additional survey projects such as this. While we recommend the future use of online surveys for collecting job analysis information and believe there remains potential for building on the AUTOGEN software to do this, there remains considerable work to adapt current procedures, tools, and staffing levels to the occupational analysis requirements to support a high stakes testing program.

Test Development

Job Knowledge Tests

Most people are familiar with multiple-choice test items. Examinees are presented with an item stem, usually in the form of a question, and asked to select the correct answer from the, typically, three to five options listed. With multiple-choice items there should be only one correct answer. Many people are also familiar with other types of job knowledge items such as matching (e.g., match the following respiratory volumes with their correct approximations), ranking (e.g., place the following steps to clear an M16 in the correct order), or drag and drop (e.g., use the computer mouse to correctly label the parts of a shuntwound generator). Computer-based testing makes developing these non-traditional job knowledge items attractive because they are an efficient and novel way to present test content to examinees. As discussed in Chapter 3, Soldiers felt the non-traditional items were a welcome change from the standard multiple-choice items they typically see.

The content of a job knowledge test is based on a job analysis and follows these steps: (a) prepare test blueprint, (b) develop items, (c) pilot test, and (d) develop final test form. Prototype computer-based job knowledge tests were developed for the 19K, 63B, and 68W MOS. As mentioned above, these prototype tests were not intended to provide comprehensive coverage of the tasks for each MOS. As with the Army-wide pilot assessment, when developing items for the MOS tests, we took advantage of the technology and used as many graphics and developed as many non-traditional items as possible.

Prepare Test Blueprint

Test blueprints are normally based on a job analysis and specify (a) the total number of items to be on the test, (b) the content areas that the test will cover, (c) the number of items to be in each content area, and (d) the organization of feedback (if any) provided. They may be based on tasks, competencies, skills, or knowledges. The 19K MOS had a recently completed blueprint

for the Select21 project (Knapp, 2003). So, rather than repeat the process, the blueprint from the Select21 project was used.

For the 68W MOS, a test blueprint was developed using the applicable STP. An STP contains the critical tasks for an MOS along with performance steps and rudimentary evaluation criteria. The table of contents organizes the tasks into subject areas. Figure 8 contains a small portion of the table of contents from STP 8-91W15-SM-TG for the 68W MOS.

Skill Level 1		
Subject Area 1: Vital Signs		
081-831-0010	MEASURE A PATIENT'S RESPIRATIONS	3-1
081-831-0011	MEASURE A PATIENT'S PULSE	3-3
081-831-0012	MEASURE A PATIENT'S BLOOD PRESSURE	3-6
081-831-0013	MEASURE A PATIENT'S TEMPERATURE	3-9
081-833-0164	MEASURE A PATIENT'S PULSE OXYGEN SATURATION	3-12
Subject Area 2: Emergency Medical Treatment		
081-831-0018	OPEN THE AIRWAY	3-14
081-831-0019	CLEAR AN UPPER AIRWAY OBSTRUCTION	3-16
081-831-0046	ADMINISTER EXTERNAL CHEST COMPRESSIONS	3-19
081-831-0048	PERFORM RESCUE BREATHING	3-23
081-833-0161	CONTROL BLEEDING	3-27
081-833-0167	PLACE A PATIENT ON A CARDIAC MONITOR	3-30
081-833-3027	MANAGE CARDIAC ARREST USING AED	3-32

Figure 8. Excerpt from the 68W STP Table of Contents.

The 68W MOS STP includes 17 task areas. SMEs were asked to weight each area so that the numbers assigned totaled 100. This tells us what percentage of test items should cover each task area. Then, for each task area the SMEs were asked to rank order the tasks in order of importance (1 = most important).

As has been mentioned, the Army tends to define jobs in terms of tasks. We wanted to explore developing competency-based assessments for the 63B and 68W MOS. For the 68W MOS, we used the results of the task-based blueprint to guide us in developing competency-based test content. That is, the results suggested that Emergency Medical Treatment and Respiratory Dysfunction/Airway Management were the two most important task areas. Using the AIT training modules and the *Brady Emergency Care* book, we developed items that tapped the competencies underlying those most important task areas. Additionally, we relied heavily upon guidance from our SMEs who suggested that, for the most part, these competencies were the anatomy and physiology of the airway and circulatory systems. Therefore, our competency-based items covered those areas.

For the 63B MOS, we completed a competency-based exercise, as opposed to a task-based exercise. This was accomplished using TM 9-8000, *Principles of Automotive Vehicles*, which is a manual that underlies what is taught at AIT. The process was very similar to the task-based blueprint exercise described above using the STPs. The manual is arranged with mega-competency areas containing underlying competency areas, which, in turn, contain individual competencies. So, as above, we asked the SMEs to weight both the mega-competency and

competency areas. Then, they ranked the individual competencies. Based on these results, we focused our efforts on engines and electrical systems. Figure 9 shows a portion of the table of contents from TM 9-8000.

CHAPTER 5	DIESEL FUEL SYSTEMS.....	5-1
Section I.	Characteristics of Diesel Fuels	5-1
Section II.	Combustion Chamber Design.....	5-3
Section III.	Injection Systems	5-8
Section IV.	Fuel Supply Pumps	5-28
Section V.	Governors.....	5-30
Section VI.	Timing Device.....	5-33
Section VII.	Cold Weather Starting Aids	5-35
Section VIII.	Fuel Filters.....	5-37
Section IX.	Engine Retarder System.....	5-39
CHAPTER 6	PROPANE FUEL SYSTEMS.....	6-1
Section I.	Characteristics.....	6-1
Section II.	Basic System.....	6-1
CHAPTER 7	EXHAUST AND EMISSION CONTROL SYSTEMS	7-1
Section I.	Exhaust System.....	7-1
Section II.	Emission Control System.....	7-2

Figure 9. Excerpt from the TM 9-8000 Table of Contents.

Develop Item Content

Developing job knowledge items is an iterative process. This process can begin with items being developed by SMEs with training provided by test developers or by item developers using appropriate reference material. For this project, the items were mostly developed by project staff using various technical manuals (TMs), field manuals (FMs), and other training material (e.g., AIT training modules for the 68W MOS).

Item development is not complete without review by SMEs. Items are often reviewed and revised several times by different SMEs, which is important to ensuring the items are clearly written and appropriate for the testing purpose. The test items developed for the two MOS for which we developed job knowledge items (63B and 68W) were reviewed by project staff and Army SMEs. The factors considered when reviewing the items are listed below:

- Is the wording of stem and response options appropriate for our target test-taking population?
- Is the keyed option correct and the unkeyed options incorrect?
- Is the item content still applicable or current?
- Does the item belong in the blueprint category in which it is placed?
- Is the item too trivial?

A further step is to collect content validity ratings, which assess the items' relevance and criticality to the job. For an operational assessment it would be best to collect these ratings from three to six judges (more for heterogeneous MOS) who have not been involved in the item

development process. Figure 10 shows the questions typically asked to gather these ratings. We did not collect content validity ratings for the items we developed.

How important is the knowledge or skill required to answer this item for acceptable performance?	
(1)	Not important
(2)	Minimally important
(3)	Moderately important
(4)	Very important
Lack of the knowledge or skill required to answer this item could result in performance errors that might cause:	
(1)	no negative consequences.
(2)	minimally damaging consequences.
(3)	moderately damaging consequences.
(4)	seriously damaging consequences.

Figure 10. Content validation ratings questions.

Pilot Test

We plan to administer the draft test items to E4 Soldiers in the Phase III pilot test. Based on the results of the pilot test we will be able to determine if the items (a) should be deleted from the bank, (b) need further revision, or (c) are ready for operational use. The result will be an item bank with associated statistics that can be used as the foundation for any operational assessment developed. As explained in Chapter 2, pilot testing of knowledge items in an operational program would be accomplished by "seeding" new experimental items in the operational test.

Form Development

For purposes of this research, we will not develop final test forms, but rather keep all items that perform well in the pilot test. In an operational setting, it will be necessary to develop a strategy for constructing multiple equivalent forms of each job knowledge test. Most likely IRT will be the best tool for this (see related discussion in Chapter 2).

Summary Comments

Job knowledge tests provide an efficient method for assessing knowledge related to important job requirements. They allow for sampling Soldiers' knowledge of all important aspects of their MOS (i.e., testing can be quite comprehensive) with items that are relatively easy and inexpensive to develop and administer. Consider that a 2-hour examination could be designed to cover all critical components of an MOS whereas only a small number of critical job tasks could be covered in a similar time period with hands-on tests or high fidelity simulations. Computer-based delivery allows inclusion of "enhanced" non-traditional items (e.g., drag and drop) that make the testing experience more appealing and realistic to Soldiers and may improve the accuracy of the

assessment as well (see Chapter 3). Job knowledge tests also have the advantage of allowing for useful feedback for examinees. Soldiers can be given a profile that shows how they scored on major portions of the test. Finally, Soldiers participating in the pilot of the core knowledge test described in Chapter 3 thought this type of test assessed their knowledge well.

As discussed later in this chapter, a number of the Army SMEs participating in PerformM21 test development work were concerned that it would be difficult to develop tests that would be fair to all Soldiers within an MOS, given differences in assignments and training. Although we did not do this in PerformM21 and we are reluctant to suggest its use in a high stakes assessment, it is possible to design knowledge tests to include item tracking or modules. Items can be tracked, in the sense that equivalent items are written for Soldiers in different positions. For example, the test could include questions related to maintenance of personal weapons that has items written for each different type of personal weapon. A test module differs from tracking, in that there is no equivalent set of items. For example, the Select21 11B test has a module of test questions on Bradley Fighting Vehicles. Only 11B Soldiers who have been trained on the Bradley equipment are administered these questions. Item tracking and test modules, however, could potentially introduce perceptions of unfairness that are worse than the fairness concerns that prompted their use in the first place. The system would also have to prevent gaming where Soldiers take those modules or tracks they will do the best on rather than the ones for which they should be responsible. If used, this is an option that must be applied judiciously.

There is a great deal of research and experience with high stakes, large volume job knowledge testing programs that the Army can draw upon to create high quality assessment programs for each MOS that have such tests. Even the complicated psychometrics associated with scoring non-traditional items and handling test modules and item tracking have a good (and steadily improving) research and experience base.

Situational Judgment Tests

Situational judgment tests (SJTs) require examinees to evaluate alternative actions to problem scenarios (Motowidlo, Dunnette, & Carter, 1990). An SJT item presents a problem situation and several possible actions to take in each situation. The problems and actions are typically presented in text form, but may be presented via video of actors or using animated characters (Weekley & Jones, 1997, 1999). Examinees may evaluate the actions in several ways, such as by rating each on an effectiveness scale or by selecting the most effective and/or least effective options. SJTs are usually scored using expert judgments provided by SMEs. The realism of SJTs comes from deriving problem situations (scenarios) and alternative actions (response options) based on what actually happens on the job (e.g., using critical incidents job analysis).

Development of an SJT typically involves the following steps: (a) identify target performance areas (e.g., customer service, conflict management), (b) develop item content, (c) develop response format and scoring key, (d) conduct a pilot test, and (e) develop final test form. An SJT designed for E4 and E5 Soldiers (regardless of MOS) for use in promotion decisions was developed as part of the NCO21 project (Knapp et al., 2002). In validation research, performance on this instrument was strongly associated with other performance measures (in particular, supervisor ratings). A short version of the LeadEx was included in the Army-wide assessment

pilot test described in Chapter 3. We developed two new prototype SJTs with content specific to two of the PerformM21 MOS – 31B and 68W.

Identify Target Performance Areas

The first step in developing an SJT is to identify the target performance areas of interest. Target performance information was collected during the 31B and 68W proponent site visits. This information was collected from NCOs, though it would have been possible collect data from the incumbent population if they had been available to us.

For the 68Ws, our review and ratings of the GWAs during the initial site visits suggested that skills with an interpersonal component were fairly important in this MOS, but not included on the task list (e.g., Contributing to and Supporting Teams; Communicating with Supervisors, Peers, and Subordinates; and Establishing and Maintaining Interpersonal Relationships). Interpersonal skills are better suited for testing by an SJT than a job knowledge test. Consequently, we decided to create an SJT that measured interpersonal skills. Similarly for the 31Bs, the SMEs at the initial site visits expressed concern that technically proficient Soldiers were being promoted who lacked interpersonal skills. Therefore, in addition to developing SJT items targeting the core 31B functions (Maneuver and Mobility Support, Police Intelligence Operations, Law and Order, Area Security, Internment, and Resettlement), we developed SJT items that measured interpersonal skills.

Develop Item Content

Development of an SJT requires a special step in the job analysis phase to generate critical incidents. Critical incidents (CIs) are actual examples of particularly effective or ineffective performance that become the scenarios for the SJT questions.

During 68W CI development, we requested that the SMEs focus on content with an interpersonal component. However, this proved to be a difficult task for them. Once we lifted the restriction, they were fairly prolific, developing 50 incidents during one site visit as shown in Table 25. Note that what they did develop reflects the results from the task list blueprint and GWAs (see Appendix D). That is, Emergency Medical Care and Interpersonal Skills are important in this MOS.

Table 25. Critical Incident Categories for 68W

Competency Area	Number of Scenarios
Communication	13
Emergency Medical Care	13
Triage	11
Preventive Maintenance Checks and Services (PMCS)	6
Planning	5
Teamwork	1
Continuing Education/Professional Development	1
Total	50

The 31B SMEs were also instructed to focus on generating CIs targeting interpersonal skills. However, as was the case with 68W, the SMEs struggled with generating examples of interpersonal CIs. On the other hand, they had little difficulty generating CIs related to four of the five core 31B functions. The SMEs informed us that 31Bs do very little related to Police Intelligence Operations, as this is the primary role of 31E Soldiers. Therefore, the SMEs did not focus on generating CIs for this function.

It may be that SMEs in both MOS struggled with generating interpersonal CIs because interpersonal behaviors are more abstract than the core MOS functions, which tend to be more tangible and established. Another contributing factor may be that we lacked good job analysis information due to limited resources. That is, the job analysis did not sufficiently identify specific KSAs, critical to performance in each MOS, on which SMEs could base their CIs.

Once the CIs were collected from the initial site visits, project staff edited them into SJT scenarios. At subsequent site visits, another group of NCOs provided feedback on the scenarios and generated response options for the scenarios. In most cases, a different group of SMEs was used at each site visit. However, in some cases, the same SMEs participated in multiple site visit workshops; this is not a problem as long as many different SMEs have input into the items. Finally, HumRRO staff edited the SJT items for grammar, accuracy, realism, richness, and clarity. After final editing, there were 33 and 30 SJT items, respectively, for the 68W and 31B MOS.

Collect Effectiveness Ratings

In order to develop a scoring key, SME ratings of the effectiveness of each response option were needed. A slightly different process was used to collect effectiveness ratings for each of the two MOS.

The original plan was to finalize the items and collect effectiveness ratings at the third and final site visit. However, due to limited resources this was not possible for the 68W MOS. Additional editing and response option generation was needed. At the third site visit to Fort Sam Houston, the SMEs were given paper copies of the SJT items and instructed to provide (a) feedback on the scenarios, (b) feedback on the existing response options, (c) effectiveness ratings for the existing response options, (d) additional response options, and (e) effectiveness ratings for the newly generated response options. The SMEs worked on the SJT items either individually or in small groups. HumRRO staff consolidated the feedback and revised the items for content and style.

Due to substantial revisions that occurred as a result of the third site visit, we needed to confer with the SMEs one more time to review the items. We emailed copies of the SJT items to the SMEs and handled the review in a teleconference. Following the teleconference, the SJT items were formatted to collect effectiveness ratings from 17 SMEs who were identified by the 68W proponent POC.

Final editing of the 31B scenarios and response options was accomplished at the third site visit. The items were displayed via a liquid crystal display (LCD) projector for the group to discuss and revise (as needed). Since revisions were finalized during the site visit, effectiveness ratings were collected at that time as well. With SJT items, as with job knowledge items, small

group (as opposed to individual) review and revision is desirable because it is efficient and generally results in higher quality items. For example, the group discussion process helps SMEs look at items more critically and brainstorm improvements.

Select Response Options and Items

Decisions about which items and response options to retain for the pilot test in Phase III were based on an examination of the item-total correlations, distribution of rating frequencies, and standard deviations (*SDs*) of the mean effectiveness ratings. One complexity related to the third 31B site visit was that six instructors attended the first day of the workshop, and 16 BNCOC students attended the second day. While individuals within each group provided independent effectiveness ratings, only the AIT instructors provided feedback on the scenarios and response options. Due to the differences between the two groups, we examined the instructor and student ratings separately; consequently, two extra steps of analyses were conducted for 31B that were not necessary for 68W.

To determine whether any SMEs tended to order the relative effectiveness of response options (for a given SJT item) in a manner inconsistent with other SMEs, we examined whether any raters consistently had low correlations with the mean effectiveness ratings of the other raters in that group. We did this by examining the number of times the internal consistency reliability (Cronbach's alpha) increased (across items) if that rater was removed. For the 31B SJT, none of the raters were consistently out of line; consequently, all raters in both the instructor and student groups were included in the analyses. One of the 68W raters was consistently out of line with the other raters; removing that rater increased coefficient alpha for 31 of the 33 items. Consequently, this rater was eliminated. Moreover, the coefficient alpha for one of the 68W SJT items was 0.073, indicating poor internal consistency among raters. Therefore, this SJT item was removed.

Next, we examined the *SDs* of the mean effectiveness ratings for each response option. Because the 31B instructors discussed the items as a group before providing independent effectiveness ratings, we expected higher agreement among the instructors; therefore, we assigned a more stringent *SD* criterion to the instructor group. If the *SD* for the instructor group was 1.75 or higher, then the response option was flagged as having low agreement. If the *SD* for the student group was 2.0 or higher, then the response option was flagged as having low agreement. We also used the 2.0 or higher decision rule for flagging the 68W ratings.

For the flagged response options, we examined the frequency of the number of times SMEs gave the option a particular rating (e.g., 1, 2, 3). If the high *SD* was due to an outlier (e.g., one rater assigned a rating of "2" and the others all gave it a "6" or higher), then that response option was retained. Outliers were not dropped; however, this will be considered when developing the scoring key (e.g., drop the ratings of the outlier SME and then recompute the mean effectiveness rating).

For the 31B SJT we also examined the mean difference between student and instructor ratings for each response option. If the mean difference between the instructor and the student ratings was 2.0 or greater for one-third or more of an item's response options, then that item was eliminated (this was seen as denoting major disagreements among the rater groups). It is important to note that an accepted strategy for identifying good SJT items is to look for items where the experts have high agreement and the target examinees have low agreement. While the

BNCOC students more closely resemble the target examinees than the AIT instructors, these students were two or more pay grades higher than the target examinees, and therefore considered experts relative to the target examinees.

We revisited the 31B SJT response options that had been flagged as a result of the *SD* for the student group being 2.0 or higher. For those cases, we looked at the *SD* of that same response option for the instructor group. If the *SD* for the instructor group was 1.0 or less and if the frequency table of instructor ratings indicated strong agreement, then we retained the response option; this only impacted 5% of the response options. Because the instructors are considered our preferred SMEs, in this way their ratings were given somewhat more weight.

In sum, this multi-step process resulted in the elimination of 3 of the 30 31B SJT items, and approximately 20% of the original response options. The process resulted in the elimination of 9 of the 33 68W SJT items and approximately 19% of the response options.

Response Format and Answer Key

It is our recommendation that in an operational application, respondents taking an SJT would select the most and least effective response options, rather than providing effectiveness ratings for each response option. Research has shown that when examinees' effectiveness ratings are compared with the mean expert ratings, those examinees assigning a rating of "4" to every response option (on a 7-point scale) score approximately 0.70 *SDs* higher than examinees seriously taking the test (Cullen, Sackett, & Lievens, 2004). Moreover, providing effectiveness ratings for all of the response options requires more administration time. For these reasons, examinees taking an operational test would be instructed to select the most and least effective options. However, for the pilot test in Phase III we will continue to ask participants to provide effectiveness ratings for all of the response options. Rating the effectiveness of all the options will help us to finalize the items and provide more flexibility in selecting the best scoring algorithm. We will also keep all of the response options that passed the above review in case some of the items that worked well in Phase II do not work well in Phase III pilot testing. This means that the number of response options will vary by item.

Pilot Test

The draft test items will be administered via computers to E4 Soldiers in the Phase III pilot test. Using those data, we will finalize the test items and scoring key. In an operational setting, it would be desirable to embed pilot items on operational test forms as would be done with the job knowledge tests. A complication is the use of different response formats (i.e., rating the effectiveness of all options versus identifying the most and least effective options). Depending on the other alternatives that might be considered, this complication may be significant enough to decide it would be best to always rate the effectiveness of all response options despite the increased administration time and related scoring issues.

Form Development

For this research, we will not develop final test forms, but rather keep all SJT items that work well during pilot testing. In an operational setting, it will be necessary to develop a strategy for constructing multiple equivalent forms of each SJT.

Summary Comments

SJTs offer a better means of measuring higher-level judgment than traditional job knowledge-based multiple-choice tests, but still offer many of the advantages of knowledge tests (e.g., group administration, computer scoring). Unlike knowledge tests, however, it is generally not possible to provide SJT score feedback that would be useful for development. Most SJTs, including those developed here, result in a single overall score.

Soldiers participating in pilot testing of the Army-wide SJT (the LeadEx), as described in Chapter 3, were comfortable with this type of assessment. The challenge is to figure out how to maintain an operational SJT program since no "industry standard" for the development of alternate forms and item seeding currently exists.

Adapting Existing Simulators

There is considerable appeal to the idea of having dual-use technology as a means of making high technology testing and training options more cost-effective. The Army has invested a great deal of resources into computer-based simulators that are used for training Soldiers. For four of our target MOS, we explored the possibility of using training simulators for testing purposes.

14E Training Simulators

We explored the viability of building assessments to run on two Patriot simulators used to train 14E Soldiers. The Patriot Organizational Maintenance Trainer (POMT) is considered a simulator. It is old equipment housed at Fort Bliss used to train 14E maintainer Soldiers on technical equipment repair and assembly. The POMT could be used as part of a hands-on assessment, but as described earlier, this methodology is resource intensive.

The Patriot Conduct of Fire Trainer (PCOFT) is a high technology simulator that supports training 14E operators for tactical air battles. The PCOFT can be programmed to test different situations, and it can be halted during a scenario for instructor interventions. The PCOFT would require additional programming to provide more measures or adaptation to testing mode, which would likely be very expensive.

Neither simulator is particularly adaptable to promotion testing for several reasons:

- Both were developed for training and collect limited performance data. For example, the PCOFT includes only three, very general measures (e.g., percent of assets defended).
- The simulators are located only at Fort Bliss so Soldiers stationed in other locations cannot be assessed.
- Each simulator assesses only one aspect of the MOS.
- Both were designed for training and many important elements in a real system are not in the simulator(s) (e.g., in the PCOFT, there are no unexpected conditions, scenarios do not model realistic flight paths, and there is no communication to enable complex problem solving).

- There is not sufficient equipment to meet current training needs, and additional uses would further exacerbate severe constraints in equipment availability.

Therefore, we decided to develop a stand-alone simulation that can be accessed on readily available personal computers at any location. This simulation is described later in this chapter.

19K Simulators

The Armor community (MOS 19K) has long been a leader in the development and use of simulators in training. Available simulators and systems cover a full range of sophistication. At the highest fidelity end are the full-crew, interactive, virtual, full-environment simulators such as the Close Combat Tactical Trainer (CCTT) which is a network of manned, collective training systems replicating heavy vehicles (tanks and infantry fighting vehicles), computer generated forces (both friendly and enemy) connected by local area network ethernet and operating together on a synthetic, virtual battlefield. Not far behind in sophistication are crew training simulator units, such as the conduct of fire trainers (COFT) used for gunnery training for commanders and gunners. These also offer a high fidelity equipment look and feel, replication of battlefield and environmental conditions, and equipment operating conditions such as ballistic solutions, ranging, and environmental degradation. There are also a number of part-task trainers, including devices that replicate the Force XXI Battle Command Brigade and Below (FBCB2) and the displays and capabilities of the digital systems found in the M1A2/M1A2SEP tanks (Crew Station Trainer – CST). Most of these trainers are widely available in both CONUS and OCONUS, including mobile versions that are deployable.

However, ultimately we were not able to fully pursue the use of these simulators to support the PerformM21 19K test development effort. Most of the more sophisticated full fidelity simulators are full crew collective tactical simulators and do not readily support many of the individual automotive and loader related tasks found at SL1. Even the gunnery simulators (COFT) primarily involve SL2 tasks and require gunner-commander interactions rather than individual actions. And although simulators have a sufficient basis of issue (BOI) to support most unit collective training and gunnery programs on a scheduled basis, they are not in sufficient numbers or in appropriate locations to adequately support a full-scale individual test program. Finally, from a developmental effort, we would have required a great deal of access to the simulators and to SMEs to adapt their inherent collective performance features to an individual test application; a level of support that exceeded the assets of the current project. However, we still believe there is much potential in the use of existing simulators in individual 19K testing. Their high fidelity depictions, economy of operation, and controlled conditions as well as their ability to electronically capture performance data are reasons enough to sustain interest. We encourage the proponent's continued investigation of this resource as part of a longer-term development program.

31B Training Simulators

The Engagement Skills Trainer (EST) 2000 is a virtual weapons training system used to train tactical skills and prepare Soldiers for battle. The EST simulator provides marksmanship, collective, and judgmental shoot-don't shoot (SDS) training. Based on input from SMEs, we believe that the marksmanship and SDS components of the EST have potential for 31B

competency assessment. The marksmanship component provides objective data, including weapon angle, trigger pressure, and shooting accuracy, that could be used to evaluate Soldiers' marksmanship skills. In fact, the EST has an M9 qualification test that parallels the actual MP qualification test.

The SDS component of the EST includes video-based situations that require Soldiers to interact with characters and determine if and when to shoot. There are 15 MP-specific SDS scenarios, which last about 1-2 minutes and include 8-12 possible outcomes that trainers control based on how Soldiers handle each situation. Unfortunately, there is no objective way to measure performance on the SDS scenarios. Rather, trainers provide Soldiers verbal, qualitative feedback at the end of each scenario. Given this, we developed ratings scales with behavioral anchors that will allow for a more objective evaluation of performance on this component of the EST.

We worked with SMEs (i.e., primarily 31B school trainers who operate the EST) to determine (a) the performance dimensions the SDS scenarios could be used to evaluate and (b) observable behaviors that exemplify low, moderate, and high performance on each dimension. We identified five dimensions that appear to be measurable with the SDS scenarios—communication skills, judgment, reaction time, marksmanship, and technique. The rating scales developed to assess each dimension are described below.

- The Communication Skills scale is intended to measure the extent to which MPs use good communications skills to control suspects during the SDS scenarios. Specific anchors describe the manner in which Soldiers identified themselves as a police officer, issued correct commands, and communicated with the suspect (e.g., confidently, clearly).
- The Judgment scale is intended to measure whether Soldiers fired their weapon at the appropriate time (e.g., "Fired only when there was a clear threat").
- The Reaction Time scale is intended to measure the speed with which Soldiers respond to the situation. That is, how long do they take to return fire or issue commands to the suspect.
- The Marksmanship scale is intended to measure Soldiers' marksmanship skills. The scale was developed to measure shooting accuracy and the number of shots fired. Example anchors on the low, moderate, and high ends include, "Shot excessively at suspects," "Fired more shots than necessary to neutralize suspects," and "Fired one or two shots at target."
- The Technique scale is intended to measure the extent to which Soldiers use proper techniques while firing a weapon (e.g., "Assumed the low ready position every time the situation escalated").

We plan to test Soldiers on four or five MP-related SDS scenarios at a time. At the end of the last scenario, two raters will evaluate the Soldiers' performance across all scenarios using the rating scales we developed. Our pilot test will use existing scenarios used for training. However, if the EST were used for promotion testing, new scenarios used solely for testing would need to be developed and programmed. Moreover, we will conduct the pilot test at a single location

given the demand for trained scorers to participate in test administration. This demand is easily resolved for our immediate purposes, but as with more traditional hands-on tests, the need for multiple scorers and essentially one-on-one testing would present significant hurdles in an operational test environment.

68W Training Simulators

Currently, the computer-based MicroSim® program is part of AIT training for medics. This PC-based program presents interactive scenarios involving patient emergencies. For instance, the Soldier may be presented with an unconscious patient and two medics who just arrived to the scene. What the Soldier will see is a picture with real people. The Soldier can then click on various parts of the body of the unconscious patient (e.g., click on the throat area to check the airway) or navigate using one of the menus at the bottom of the page to help him investigate. If the Soldier opts to check the airway, one of the medics in the scene will do as instructed and respond accordingly. Based on the response, the Soldier decides what to do next.

MicroSim® can be easily configured to test Soldiers. However, this could be an expensive endeavor and so we did not pursue it in our research. Developing such sophisticated computer-based simulations requires enormous resources – money, time, and SMEs. New scenarios would need to be developed for testing, and may need to be updated frequently. MicroSim® is an excellent tool for testing critical thinking and declarative knowledge, but not for the haptic skills, which are extremely important for this MOS. It is one thing to say a patient needs a chest tube inserted, but quite another to be able to actually insert one.

The SimMan® patient simulator has an airway system that allows for a realistic simulation of difficult airway management and patient care situations, the realistic practice of chest tube insertion, physiologically correct carotid, femoral, brachial, and radial pulses, a library of cardiac rhythm variants allowing cardiac monitoring, defibrillation, and external pacing, and a library of heart, lung, bowel, and vocal sounds, among other features. It can be programmed with a certain illness/injury via computer, or operated via remote control. Figure 11 shows one such simulator at the Department of Combat Medic Training (DCMT) at Fort Sam Houston set up in a simulated combat hospital.

Because these simulators can be programmed and they present a standardized situation to Soldiers, they can make excellent test equipment. The Army currently has approximately 400 of these simulators with 166 at Fort Sam Houston, TX. Most major posts have one or two (CPT Garrett, personal communication, 9 September 2004). One issue concerning using these simulators for testing revolve around whether they are available on a given post, and, if so, whether anyone knows how to operate them. If a SimMan® is not available on post, then one or more would need to be shipped. They weigh approximately 180 pounds, but that could double depending upon how the simulator would need to be set up for testing. Additionally, they must be shipped as freight, and when they are shipped back for repair, it typically costs more than \$300 (CPT Garrett, personal communication, 4 September 2004). There also would need to be someone who could program the simulator to run the test.

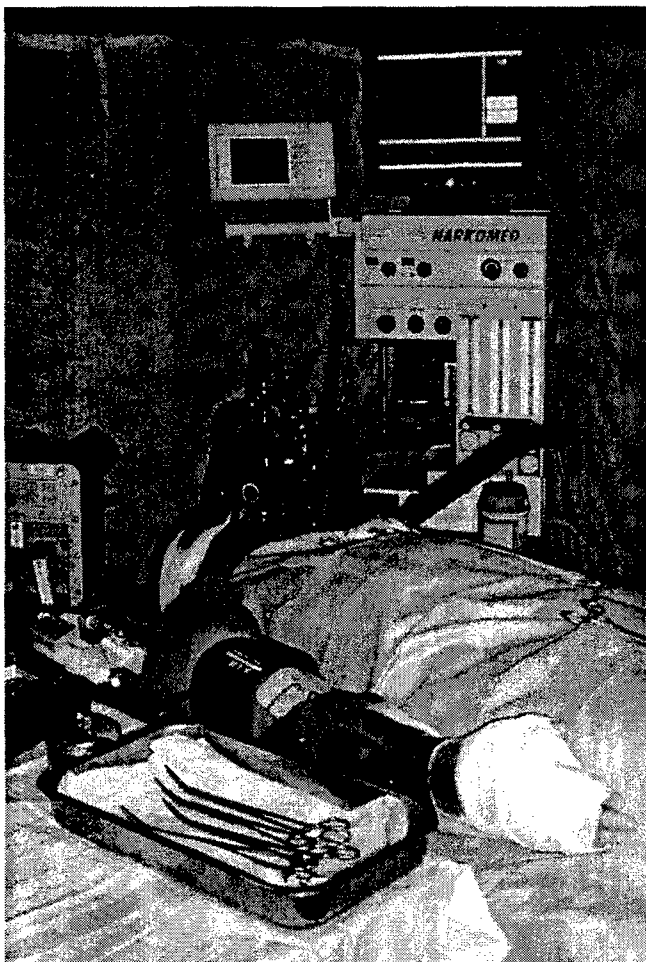


Figure 11. SimMan® at Fort Sam Houston.

A more compelling issue, however, is that a SimMan® would be useful for helping to standardize a hands-on test, but could not be used as a stand-alone assessment since there is no way for the simulator by itself to present scenario details and record Soldier performance.

Summary Comments

We expect our experience in PerformM21 is illustrative of what will happen with other MOS. That is, it will be difficult to identify training simulators that can be used for high stakes testing without considerable additional investment in the technology enhancements (e.g., to create additional scenarios, to program the software to capture performance information, to purchase more simulators). That said, the potential for dual-use technology is very real and should be a standard question for each MOS testing program. A caution, however, is that adapting simulators to serve as testing vehicles needs to be done in a manner that does not compromise their utility for training.

Developing New Simulations

Computer-based simulations complement traditional tests when there is a need to evaluate cognitively complex skills such as multi-tasking, vigilance, adaptability, and anticipating and solving complex problems in dynamic, time-critical, and information-rich environments. Additionally, simulations are uniquely suited to assessing behaviors when the consequences for incorrect actions may result in failure of critical and expensive equipment and/or loss of life.

Developing a computer simulation for an assessment involves the following activities: (a) gather simulation requirements and identify the critical performance areas that cannot be effectively assessed through traditional methods; (b) develop a description of the environment to be simulated and a set of scenarios that target the identified performance areas; (c) develop story boards describing each scenario in terms of events such as user interaction, visual display changes, sounds, and user navigation; (d) design a simulation interface; (e) develop graphics, sounds and other environmental features; (f) develop the simulation software; (g) conduct user acceptance testing and make revisions; and (h) pilot test.

The remainder of this section illustrates development of a fairly sophisticated simulation for the 14E MOS. We should note, however, that we also developed two very simple simulations for the 19K MOS using animation developed for training applications. These are point and click, path simulations (see below for an explanation of these features). One simulation takes the Soldier through the 11 steps of a .50 caliber machine gun function check and the other simulates a similar process on the M1A2SEP tank. We also developed a series of four multiple-choice items that use animation to illustrate answers to questions regarding how to handle a cold .50 caliber machine gun (i.e., one that will not fire).

Gather Simulation Requirements

Project parameters and SME input guided our simulation requirements in PerformM21. The project was designed to create a “proof of concept” simulation to determine whether simulations are an appropriate methodology to evaluate MOS proficiency and readiness for promotion in a high tech MOS. The project scope, budget, and timeline limited the primary simulation scenario to a small segment of the 14E Soldier’s job.

We conducted two workshops to obtain 14E SME input into the simulation requirements. The SMEs completed surveys, reviewed task lists, and created a rough test blueprint using the Soldier’s Manual (STP-14E1-SM). Given the technical and cognitive complexity of the MOS, the SMEs thought that a computer-based simulation made sense as a testing methodology. They wanted the simulation to incorporate a high degree of realism and for it to assess essential skills and abilities not readily evaluated in traditional tests.

Project requirements included the need for the simulation to balance realism with affordability and technical requirements. We were mindful of striking this balance throughout the remaining steps of the simulation development process.

Develop Environment and Scenarios

In defining the scenario, the SMEs decided to evaluate the Soldier’s ability to operate the relevant equipment to troubleshoot a fault by using their manuals to follow procedures. Also, the SMEs wanted the scenario to reflect the reality that their work occurs in a team environment. We explain later how we used sounds and dialogue boxes to reflect this work feature. The design incorporates the ability to move from place to place and actually operate equipment, and provides access to the technical manuals for reference.

A key element that defined how scenarios worked from the point of view of the developer was whether the simulation would be an open simulation or a path-based simulation. A path-based simulation separates tasks into discrete, linear steps. Simulation authors create a graphical representation of each step and then link them together to recreate the path used to complete the task. Table 26 describes path-based and open simulation features and the impact on relevant requirements and potential risks.

Table 26. Features of Path and Open Simulations

Features	Requirements Satisfied	Potential Risks
<i>Path Simulation</i>		
Simulate only a specific procedure, not the entire target system	Reduces software development costs	Low realism
Predictable development time	Affordable	Changes can be expensive
<i>Open Simulation</i>		
Simulation engine is developed one time for all scenarios.	Reduces time to develop multiple scenarios	Writing one simulation engine for each MOS might not be affordable.
Simulation engine accounts for every possibility because it simulates every piece of equipment.	High realism	Determining mathematical models to simulate very complex systems is very expensive.

An open simulation recreates the target system, allowing for unscripted actions within the simulated environment. The simulation is “open” because allowed actions are relatively unrestricted. A robust open simulation environment has several key components (Whitnah, 2004):

- Simulation engine: recreates the interface and logic of the target environment.
- Scenario data: separate from the simulation engine logic, allowing easy modification of the data to correct problems or create unique simulated environments. The simulation engine pulls in data for display as the simulation runs.
- Scoring criteria: since it is external to the simulation engine, new test items or scenarios can be created by modifying the scoring criteria, not the simulation engine logic.

For this particular application, we chose a solution that mixes elements from both path-based and open simulations in order to keep a reasonable level of realism while minimizing long-term software development costs (e.g., if this assessment format is applied to other MOS). The simulation engine recreates as much from the target system as is needed for a Soldier to have the ability to complete the scenarios, but the simulation does not react to every possible action, just a sequence pre-determined by the simulation author. We created a process model to represent this sequence of actions independent of MOS-specific requirements. The model allows the software developers to focus on what makes an MOS unique when creating a new simulation-based assessment. Code that requires modifications in order to implement additional scenarios, or a simulation-based assessment for a different MOS, is isolated. These features have been implemented in frameworks that eventually, with additional research, can be turned into graphical tools for easier simulation development and maintenance, either within or across MOS.

Develop Storyboard

The next step is to create the storyboard to detail the content. The storyboard communicates the steps the Soldier is expected to take in order to solve the problem, how each screen of the simulation will appear – its text and images – and the functionality of each screen. In addition, the storyboard communicates the consequences of the possible actions a Soldier might take, whether correct or not, in terms of simulation behavior and scoring.

We met with the SMEs on web-enabled conference calls to review and finalize the storyboard. After several iterations of feedback and changes, the storyboard was approved as the basis for the prototype simulation-based assessment.

Design a Simulation Interface

There are two major types of interfaces for simulations, which result in two different levels of realism and development cost. A simulation can have a real-time interface, where the user moves around a 3-dimensional (3D) virtual world similar to the way many computer games operate. While this method provides a very high level of cognitive realism and is the method of choice for many high-end computer-based simulations, there are several disadvantages:

- Off-the-shelf and customized software engines that allow the creation and visualization of 3D virtual worlds typically cost from hundreds-of-thousands to millions of dollars. They also require additional customizations so that they can be used as effective assessment tools.
- Creating a complete 3D model of the environment including panel buttons and indicators increases the costs.
- Subsequent changes to the assessment are expensive due to the effort required to modify the 3D environment model.
- Visualizing 3D virtual worlds in real-time increases the hardware requirements of the simulation.
- Operating the simulation is cognitively complex. It requires proficiency in the navigation of a virtual 3D world. Actions such as walking, turning, climbing, and jumping require simultaneous operation of the keyboard and mouse.

For the above reasons, we evaluated using a click-through interface that is used in games featuring a photographic level of realism. Different from 3D, the user interacts with the environment by clicking on areas of the screen, each triggering a different behavior (i.e., result). The advantages of such an approach include:

- The simulated environment can be represented by a set of 2-dimensional (2D) graphics, which an artist can produce relatively inexpensively.
- The simulation itself is far easier to operate, requiring only the computer mouse.
- The scenario drives the complexity of the simulation, rather than the software and hardware requirements.

One consideration in our design was that simulations may be incorporated in assessments for other high technology MOS, and a click-through interface makes the production of these simulations more affordable. Therefore, for the 14E prototype we chose a click-through interface between the Soldier and the simulated environment.

Develop Graphics, Sounds, and Other Environmental Features

Graphics and environmental features were developed by an artist using a combination of illustration skills and photographic material obtained through several visits to Fort Bliss. Combining these resources made the process of creating a believable environment more efficient for the graphic artist. Reproductions of control panels and other equipment were reviewed and approved by the SMEs. Figure 12 shows a screen shot of the simulation.

Several audio sources were also recorded during visits to Fort Bliss, and later edited for use in the simulation. These sounds add to the realism of the simulation by incorporating background noise, audible alarms, and credible dialogue that provides Soldier feedback and reinforces the team environment.

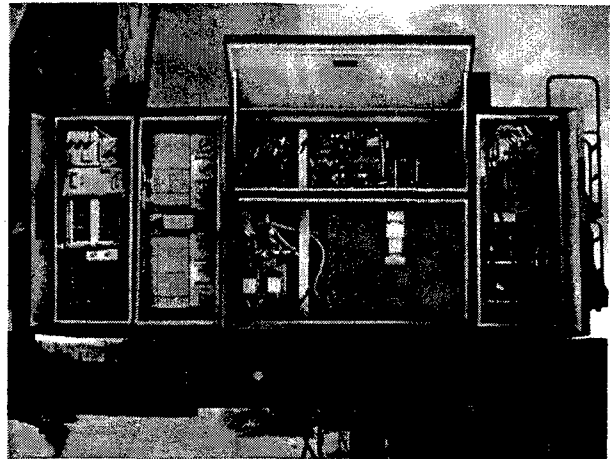


Figure 12. Screenshot of 14E simulation.

Develop the Simulation Software

With the design complete, the software can be developed. One of our objectives was to apply simulation-based assessment techniques to multiple MOS, with the possibility of reusing the same software core throughout the development of different simulations. A process-model representation of simulation behavior was designed that supports this objective. By using such a representation, the core software that runs the simulation can be used unchanged for different MOS, significantly limiting the software development effort required to create new assessments. With additional research and development, tools can be created to make it easier and more efficient to develop and modify simulation-based assessments. Additionally, the software and the same graphics can be used for other scenarios in the same assessment as well as for evaluating other levels of Soldier proficiency within the same MOS. Thus, the high up-front development costs are mitigated with additional uses.

In order to simplify development activities and ensure greater compatibility, we chose Java 2 Standard Edition as our programming platform. This choice also allows the simulation to be somewhat web-enabled, meaning it is theoretically possible to run it through the web as a Java Applet. However, there are limitations due to the multimedia nature of the simulated environment:

- The entire simulation application package must reside at the computer where a Soldier is being tested to ensure proper simulation performance.
- The size of the simulation application package is too large for it to be downloaded at the time of the test, which means that it must be downloaded prior to testing.
- Since network conditions might change while the simulation is running, it might be unable to communicate with the server-side component that will store a person's

score(s). There must be an option to manually gather the information (e.g., on a floppy disk) so that it can be incorporated into the scoring system at a later time.

Existing technologies can also assist in this matter, such as Java Web Start, but further research is needed to determine creative software deployment methods that would allow these simulation-based assessments to be launched through the web.

User Acceptance Testing

Four of our SMEs participated in an all day session reviewing and critiquing the simulation. They spent 5 hours at individual computers reviewing the simulation from an MOS technical perspective. They then worked together to prioritize their requested changes as follows:

1. Changes critical to a Soldier successfully completing the prototype simulation
2. Changes that would be nice to make, but are not essential for this prototype
3. Recommendations for the next time we develop a simulation

We made all the critical changes they requested and many of the “nice to have” modifications in category two. The following are recommendations for future simulations/enhancements.

- Currently all the switch indicators on the Engagement Control Station (ECS) console function, but only the ones essential to resolving the fault bring up tabs (data) on the monitor. It would be good to have access to some of the remaining tabs that provide additional information to Soldiers but that are not essential to resolving the fault.
- Rather than providing a pop-up window at the end advising Soldiers that the fault cleared, extend the simulation ending and return the Soldier to the ECS where, in the real world, the Soldier would verify that the fault cleared.
- Introduce a maximum time for taking the simulation. The SMEs suggested limiting the Soldier’s time to try endless options, not creating this as a “timed” test, per se.
- Add a measure indicating whether the Soldier committed any safety violations.

Pilot Testing

We will pilot test the simulation on E4 Soldiers in the Phase III pilot tests. We will be particularly interested in their reactions to this assessment method and evaluating the performance information that the simulation produces.

Summary Comments

The obvious appeal of high fidelity computer-based simulations for assessment is that they have most of the advantages of hands-on tests with fewer practical drawbacks (e.g., there is no need for trained scorers and tests can be administered to many Soldiers at once). The biggest catch with such simulations is the expense of developing them. As with hands-on tests, the method also suffers from the fact that it takes longer to assess a particular content area than it

does in something like a multiple-choice test. That is, there is a tradeoff between test time, test coverage, and the richness of the assessment experience. Another complication is that there are not yet commonly accepted practices for addressing the psychometric issues that arise when using complex assessments (e.g., how to score them and evaluate their psychometric properties). This is rapidly changing, however, so it should not prevent the Army from moving forward with these more innovative test methods. Rather, we encourage the Army to explore further ways in which the expense of programming high fidelity simulations for assessment can be offset by finding multiple uses for each software module investment.

Embedded Assessment

Evaluating actual work activities or products could offer a non-obtrusive, cost-effective means for assessing at least some aspects of job competence. After working with SMEs to identify possible candidates for this assessment method in their MOS, some of the issues that need to be successfully addressed to make the process work are as follows:

- Is the performance information important?
- Is the performance information readily accessible (e.g., computerized at a central location)?
- Is there a sensible, objective way to evaluate the performance information?
- Can scoring be done by computer or with a small set of trained raters?
- Is the performance information "uncontaminated" (i.e., a reflection only of the individual Soldier's competence)?
- Is there a good process for sampling examples of performance to ensure reliable measurement?

Military Police Report DA Form 3975

In our early meetings with SMEs, we tried to identify potentially scoreable work products for each of our five target MOS. The only MOS that had a promising candidate for this measurement method was 31B. Military Police Report DA Form 3975 is the foundation of every case handled by MPs. Form 3975 provides a record of each incident requiring police action, administrative control for complaint follow-up, and statistical data for analysis and review. It also provides unit commanders with summary and statistical information about individuals in their command. This report ensures complaints are recorded, systematically assigned follow-up, disposed, referred, and posted to the MP desk blotter. In other words, Form 3975 brings all the information together so that it can be presented in a complete format. Once DA Form 3975 is completed at the post or installation level, it is transmitted to a centralized database located at the Pentagon. This information becomes a permanent record, and is available to all law enforcement agencies worldwide to assist real time continuous crime statistics. Failure to record information properly will severely hamper and delay the effectiveness of the military police and law enforcement agencies worldwide.

Because the accurate completion of Form 3975 is a critical job task, the development of an assessment instrument based on this form is an appealing option. Another advantage is that Form 3975 is completed on the Centralized Operations Police Suite (COPS), meaning that the

information is stored into a computer database; therefore, information should be easily accessible for scoring. We obtained initial information regarding Form 3975 from STP No. 19-95B1-SM.

Exploring an Assessment Option

In training, Soldiers complete Form 3975 based on mock information provided to them by their instructors. Instructors rate the trainees' performance by providing GO/ NO-GO ratings for each section of the form. Instructors also provide a GO/NO-GO rating for the overall "completeness and accuracy" of the form. Initially, we thought about developing a standardized protocol for assessing the completeness and accuracy of operational Form 3975 reports. A sample of these reports could be scored by trained evaluators at a central site.

We identified several issues related to this idea. First, it is not possible to verify whether the information contained in a report is correct and comprehensive. In other words, there is no "answer key." Alternatively, completeness could be defined in terms of whether sections within the report were left blank, and accuracy could be defined in terms of whether the information was entered in accordance with the format specified by the STP (for example, the STP specifies the time of the incident must be recorded in military time and the date must be recorded using the DOD date format). However, these considerations raise additional issues. Form 3975 on the COPS system is formatted with mandatory fields and only accepts information entered in particular formats (e.g., it only accepts military time and DOD date formats). Moreover, completed forms are typically reviewed and edited by a supervisor so the forms are not solely a reflection of the individual's performance.

It thus became evident that the development of a rating scale to measure completeness and accuracy is problematic because there would be little or no variation in performance on these measures due to existing system checks. The 31B SMEs also felt that a rating scale to assess MP performance on Form 3975 would simply measure one's ability to follow directions and complete a form, as opposed to measuring one's investigative ability. Consequently, we did not pursue development of this prototype concept.

An Alternate Assessment Option

The problems we encountered might be addressed by using a different assessment method. Instead of evaluating actual MP forms, we could envision a relatively high fidelity, but not expensive computer-based simulation. Examinees might view video of a fairly complex scene and then complete a computerized mock-up of Form 3975 that does not include all of the quality-control features of the actual COPS-based form. In this manner, the accuracy, completeness, and format of an individual's Form 3975 report could be evaluated. In all likelihood, the scoring could also be completely automated.

Although the idea of embedded assessment was not fulfilled during this development and we suspect that it will not often be a realistic option, posing the question is useful. Searching for possibilities of embedded assessments led to other good ideas of other assessment options. Also, the payoff of identifying possibilities that do work in the future would be high.

Hands-On Tests

Hands-on tests are characterized by performance of critical parts of the job using actual equipment under standardized conditions. Performance is usually evaluated by one or more trained observers using a standardized protocol.

A Synopsis of Hands-On Testing in the Army

The Army historically has strong links to hands-on performance as part of its training paradigm. To start with, most Army job requirements fall into the realm of discrete, proceduralized, psychomotor tasks that lend themselves to standardized hands-on performance evaluation. The Army's approach to task definition is to list "conditions and standards" in which the "standards" most often are expressed in terms of the steps or performances required for task completion. Most training situations are "hands-on" in that they involve the actual equipment used in job performance. Instructional training usually follows a design where individuals (or groups) conduct hands-on practices and eventually must demonstrate proficiency by performing the task "to standard" without assistance. At this point, task training is usually deemed successful.

Within its Army usage, hands-on evaluation has generally quite high validity. The tasks identified are usually highly job relevant. Although not all job conditions can be readily duplicated (e.g., combat, full field, initiating cues), the use of actual equipment enhances the job match. Moreover, the performance steps are almost always a doctrinally required performance method that must also be applied on the job. Observers (trainers or evaluators) are usually NCOs who are experienced in the tasks being performed. Most importantly, hands-on performance has high acceptability in the Army community, particularly in the NCO corps and among most Soldiers as well. The widespread perception is that hands-on performance is the "best" way of assessing Soldier task mastery, an opinion voiced by many of the SMEs we worked with during the course of prototype test development (see the "Focus on Individual MOS" section later in this chapter).

Although Army hands-on evaluations are widely used in specific institutional training situations (e.g., initial entry training, advanced MOS training, Primary Leadership Development Course) and are also widely found in informal, localized unit training situations, their use in formalized, non-institutional applications is somewhat limited. There currently are three primary areas of this latter category of hands-on evaluations:

- The Common Task Test (CTT) – An annual requirement for all Soldiers. The test consists of 18 pre-selected tasks from the Soldiers Manual of Common Tasks, Skill Levels 1-4. The majority (12) must be "core" Skill Level 1 tasks.
- The Tank Crew Gunnery Skills Test (TCGST) – A semi-annual requirement of all tank crewmen. About 20 MOS, tank-specific tasks are tested. Similar tests exist for crewmembers on infantry, scout (cavalry), and air defense variants of the Bradley Fighting Vehicle and for Stryker and HMMWV combat vehicle crews.

- The Expert Infantryman Badge (EIB) – A voluntary, but widely administered performance test available to Soldiers in CMF 11 (and CMF 18). The test consists of approximately 62 tasks. This is the most rigidly controlled and standardized of the existing hands-on tests. A similar test (The Expert Medical Badge) exists for certain Soldiers in the medical CMF.

It would obviously be ideal to be able to use some of the Army's existing hands-on testing programs to satisfy at least some competency assessment requirements.

Hands-On Testing for Training Versus Assessment

Despite the Army's reliance on hands-on tests, it is important to realize that there are fundamental differences between hands-on *performance*, as commonly used in training applications, and hands-on *assessment* such as would be used for promotions. The high-stakes implications of hands-on tests to be used in promotion decisions magnify several problems inherent in hands-on testing.

The first issue concerns standardized administration. For training purposes, it is not necessary to ensure that all Soldiers are presented with the exact same test set-up and conditions. In comparison, highly standardized test conditions are required for high stakes assessment.

A second issue concerns scoring. Although scorer reliability is enhanced by good instrument development, scorer requirements go beyond external preparations. Scorer training and supervision must be standardized and enforced. Scorer objectivity, both actual and perceived, will always be an issue in hands-on testing and the structure needed to project and ensure objectivity in large-scale testing is cumbersome and expensive. (Of the three hands-on evaluations described previously, the EIB does the best job of ensuring test integrity, but it does so with a high administrative burden. Moreover, the EIB, being a one-time, tangible award, has some built-in tradition and checks and balances – EIB scorers and administrators must themselves be EIB holders.) Finally, the traits that make NCOs good trainers, instructors, and motivators work against them in a pure testing situation; it is often difficult to enforce pure objectivity and detachment with NCO scorers in spite of specialized scorer training.

Also with regard to scoring, hands-on tests for training do not typically require the recording of individual's scores, and if so, the only information that is recorded is whether the Soldier passed or failed. Promotion tests are considerably more informative if they provide continuous scores to reflect the degree to which the examinee has acquired the requisite competencies.

However, administration costs are usually the biggest issue in large-scale hands-on testing programs. Within the Army, the issue is seldom equipment; even major end-items (trucks, helicopters, artillery pieces) can usually be accessed for testing without undue operational burdens. But hands-on testing is done on a 1:1 basis – one scorer for each examinee. Every examinee test minute must be matched with a corresponding scorer minute. When the requirements of scorer training, site and test preparation, supervision and administration (including score sheet verification and collection), and logistical support are added in, the overhead becomes almost overwhelming.

In applications such as training, there are ways of reducing this overhead requirement and achieving administration efficiencies. Not so in large scale, "for record" applications. This was the experience of the original SQT that included a mandatory hands-on component. As the most visible part of the early SQT, it got the most attention, eventually to the detriment of the entire program. The requirement that all MOS have a hands-on component contributed to the burden. Given the mix and range of MOS within most units, commanders found themselves in cycles of never-ending support and administration that multiplied as successive MOS testing requirements were added (U.S. Government Accounting Office, 1982).

Exploring the Possibilities Further with 19K

At the start of the development of MOS-specific measures, there was interest generated in adopting or adapting the TCGST as a measure to be used as a promotion criterion for 19K. Its job relevance and performance criticality has already been established. At least on some level, it already exists as a "test." The research issue was whether the TCGST could be operationally included for promotion testing. Specifically, did it meet all the requirements of reliability and standardization as well as those of administration feasibility? We explored the specifics of the TCGST in some detail, including discussion of the concept during the workshops with 19K SMEs.

Ultimately we concluded that it would not be productive to pursue the TCGST at this time; the standardization problems were just too much to overcome given the resources available to this phase of the development. Additionally, there was some resistance on the part of the SMEs to using TCGST as promotion criteria. Nonetheless, the exercise proved beneficial for what we did discover.

Standardized administration. The TCGST allows units to determine what the test includes and how the test is administered on some stations. Station 1 (Identify Vehicles and Helicopters), is an excellent example. In this station, the Soldier has to correctly identify 18 of 20 vehicles and helicopters by name. He has to correctly identify all U.S. vehicles to pass, however, so the two vehicles he can miss must be foreign vehicles. A minimum of five U.S. vehicles and helicopters (mixed) (with no more than three helicopters per test) are selected from U.S. armored vehicles and combat helicopters currently in use. The remaining vehicles are non-U.S. vehicles. Not dictating what vehicles are shown results in varying levels of test difficulty and a varying test standard. One unit might select only two U.S. armored vehicles (e.g., M1A1, M1A2SEP) and three U.S. helicopters (e.g., Apache, Blackhawk, HIND-D), to meet the minimum requirement of five friendly vehicles and then include 15 armored vehicles from other countries. Another unit might include 20 U.S. vehicles, which, given the caveat about having to identify all U.S. vehicles correctly, means that the standard for their receiving a GO at that station is now 20 out of 20. However, the first group with more foreign vehicles may be at a disadvantage because they are less familiar with foreign vehicles than with U.S. vehicles. Similar variability is likely to occur in selecting the vehicle views and photos.

A solution to standardizing TCGST Station 1 administration might be to provide units with a computer-based assessment (of course, it would no longer be considered a hands-on test per se, but neither is the current version). TCGST Station 2, which tests the Soldier's ability to

identify 105-mm or 120-mm main gun ammunition, would also benefit from computer-based administration so that every Soldier receives the same test rather than some Soldiers being tested on mockups, others on slides, and others on photos.

Another way to standardize TCGST administration is to make sure that all stations are set up uniformly. Currently, setup requirements for the test stations are not specified. A test version would need to require test administrator setup checklists to adhere to as part of test preparation.

Another administration issue regards retesting. Currently, if a crewmember fails a task/station, he must be retrained and retested on that station until he receives a GO, but retesting procedures are in accordance with local standard operating procedure; there is nothing in FM 3-20.12 about retesting procedures. If the TCGST results were to be used for 19K MOS assessment (and even if they are not), retesting procedures should be clearly documented. As a general principal, although retesting is a standard procedure in training, immediate retesting is usually not a promotion test practice. And if it were to be, the conditions and number of retest must be strictly controlled to ensure objectivity and fairness. There may be a definite conflict in this area between the TCGST used for training and when used in a promotion application.

Scoring. The 19K SMEs contended that evaluator bias may be a problem in TCGST. And even if evaluators are objective, Soldiers being tested may perceive them as biased. There are several reasons for this possible bias. Most benignly, NCOs, are, for the most part, trainers rather than testers. They are inclined to score leniently and “give a break” as long as the Soldier can be retrained and can eventually pass the test. Complicating matters is the belief that TCGST scores are used to evaluate commanders. An understanding by evaluators that their commander’s performance record could be influenced by the TCGST results can lead to lenient scoring by evaluators who are “giving the commander a break.”

There are several possible solutions to mitigate actual and perceived evaluator bias. One option might be to have a TCGST Administration Team assigned to travel from post to post, unit to unit, overseeing the administration of the TCGST. They would be responsible for training the evaluators, making sure that the stations are set up in a standardized fashion, and providing general administrative oversight. The use of a dedicated TCGST Administration Team, however, may be too expensive.

A less expensive alternative would be to use local evaluators from outside the unit, much as is done now, but have two individual evaluators instead of only one for each station. The individual evaluators would reconcile their scores at the end, serving as a sort of “check and balance” against one another to help mitigate bias.

Score recording. The TCGST currently uses paper-based Criterion Scoring Checklists. These could be upgraded to electronic personal data assistant (PDA) checklists. Moving to computer technology would have several benefits. First, the timing element of task completion, one of the two criteria for receiving a GO at all stations, could be built in to the electronic checklist. Secondly, with electronic checklists, using wireless technology, scores could be immediately uploaded to a central database. This would offer several benefits. Immediate upload means less opportunity for scores to be changed for whatever reason. Also, electronic

scoresheets which are immediately uploaded would facilitate quicker roll-ups for the individual Soldiers, platoons, and companies to be included in the NCOIC's TCGST results briefing to the commander. Areas needing retraining and retesting could be more quickly identified. Scoring anomalies could also be more quickly noted and corrective actions more quickly taken.

Concluding Comments

Although existing Army hands-on tests, including the TCGST, are useful training tools, they would require significant modifications to satisfy individual competency assessment requirements. This is particularly true given that some of the changes needed for assessment would compromise their utility for training.

Despite some real negatives, hands-on tests for revived Army job testing should not be automatically excluded. Their high validity and acceptability alone make them worthy of consideration. Technological improvements in score sheet manifestations and recording (such as with PDAs) show promise for increasing scorer reliability and decreasing the burden of score recording, data collection, and consolidation. But their use should be worth the administration cost and the developmental efforts required to produce first-rate instruments. To this end, hands-on testing should be reserved for a few tasks in some MOS where the pay-off in Soldier evaluation is considered to be worthwhile; not every MOS needs to have a hands-on test. Development efforts should concentrate on what *should* be tested hands-on, not what *can* be tested hands-on. At least in the near term, the smaller the overall scope of the administration, the more acceptable hands-on test administration will be to those who must support that administration.

A final, but very critical consideration, is that test developers and implementers must not think of hands-on as a dichotomous choice (i.e., a test is either hands-on or it is not). In fact, performance testing is a continuum, with many options regarding job fidelity and validity. As a general rule (and particularly as applied to process-scored tasks), as the validity and the fidelity decreases, the reliability and administration efficiency increases. But it has never been an absolute choice between two extremes. Many tests which are not "pure" hands-on are treated as such by the Army (e.g., several of the TCGST tests) and are quite valid performance indicators. As simulations and computer presentations improve, the line between what is and is not a "hands-on" test becomes even less discernible. Initiative, experimentation, and imaginative thinking on the part of test developers in the varied MOS should be encouraged, rather than thinking in terms of a simple choice between whether or not to have "hands on" testing.

Operational Appraisals: Reviews of the Target MOS

In this section, we present a by-MOS review of what we experienced with each of the MOS we explored and what our appraisal is for assessment potential within the MOS. It must be kept in mind that what we did was not full-scale analysis and development. An operational test program would require much more in the way of resources – particularly time – than we were able to devote. Also, although we had excellent support and cooperation, we were not operating with the command emphasis, procedural guidance, and proponent professional resources support that would accrue to an operational program. Nonetheless, this recap gives a good idea of both the potential and problems in MOS testing.

As noted at the start of this chapter, when we made our initial site visits, we briefed the workshop participants on the concept of MOS testing and gave them a chance to react and respond. In each of the following recaps, we start off by presenting a synopsis of those responses. These are presented to provide a sense of what the field's reactions are; the statements are not presented as action requirements or even issues. Many are based on unwarranted assumptions about what a test program would look like or require and some of the statements simply do not stand up to a factual examination. Yet, they are important, because they typify what participants'—mostly E6 and higher NCOs—immediate reactions are. It is noteworthy that most immediate reactions tended to be negative. It is also noteworthy that invariably attitudes and reactions would shift later in the workshops as participation and understanding progressed.

14E Patriot Fire Control Enhanced Operator/Maintainer

This MOS was selected as representative of the types of high tech Army MOS that would be likely candidates for assessment using a computer-based simulation. Our preliminary job analysis indicated that this is a cognitively complex MOS. To be effective, Soldiers must be vigilant, avoid making premature judgments, anticipate and solve complex problems, and quickly process large amounts of information. In combination, these findings affirmed our early belief that this MOS is well suited to testing using a computer-based simulation.

SME Reactions to Testing

When asked, both sets of 14E SME workshop participants raised initial concerns about promotion testing:

- Testing Soldiers at the individual level is counter to the Army's team-based training.
- The MOS has two specialties (operators and maintainers) and if Soldiers are required to know both, then the requirements for promotion would make 14E less attractive compared to other MOS.
- The duties for every position (even within the respective specialties) are different suggesting it would be difficult to develop a fair test for everyone.
- There is no need for yet another measure since platoon sergeants only send qualified candidates to the promotion board.

After discussion, though, all the SMEs shifted to thinking that including an objective test score would be a good addition to the promotion board point system, primarily because unit commanders have different standards and the board does not evaluate MOS related knowledge or skills. They agreed that many Soldiers are promoted because they meet time-in-service and time-in-grade measures and "are not 'the worst' Soldiers" rather than promoting the most highly qualified Soldiers from among the available pool.

The Prototype Assessment

In developing the general assessment guidelines, the proponent POC indicated that the 14E MOS should be considered one occupation and that all 14E Soldiers should be assessed on

both tactical operations and maintenance. Additionally, the SMEs emphasized that any assessment should incorporate a team context and be based upon systems, not specific tasks.

Developing the prototype required balancing realism with affordability and technical requirements. We worked with the SMEs to generate a list of job requirements, which they then reviewed to identify possible simulations. The SMEs chose a simple, short scenario that evaluates a Soldier's ability to troubleshoot a fault by following procedures. This scenario met several objectives: (a) it partially evaluates operations at the Engagement Control Station (ECS) and maintenance at the Radar Set (RS) (major job requirements for the operator and maintainer, respectively), (b) it tests whether the Soldier will ask for help prior to troubleshooting (a serious problem when it fails to occur), (c) it evaluates the ability to follow procedures (which the SMEs thought would be important to test), and (d) it met our research objective of creating a scenario that showed diverse capabilities of the simulation.

Balancing realism, technology, and affordability. At the onset, the SMEs emphasized that realism was critical. We created a 2-dimensional environment that blends synthetic images with some high quality digital photographs. For example, the simulation provides synthetic images of the interior and exterior of the ECS and the RS with the ability to open and close panels and doors, and operate switches and circuit breakers on synthetic interior of the panels. Many of the synthetic images are blended with digital images to enhance the realism.

We designed the simulation to be compatible with computer systems at the two Fort Bliss DTFs (NCO Academy and the Sergeants Major Academy). In the Phase III pilot test, the simulation will be delivered on a CD and conform to current DTF technology. In principle, it is capable of being delivered over the web, but since the simulation must be installed on the computers, local support will be needed to handle any installation troubleshooting.

To address affordability we shortened the simulation to assess only the primary objective: the ability to follow procedures to solve a problem. At the same time, we also sought to incorporate a wide variety of applications that simulate the real environment such as equipment noise and functionality, maneuverability to different equipment, communications, distracters, and stopping the simulation to insert relevant multiple-choice questions.

User fidelity. Another development goal was to create a simulation with sufficient fidelity to engage the test taker while also providing intuitive computer interaction. For example, the prototype simulation provides a realistic mock-up of the operator's keyboard/console in the ECS with working keys and switches operated by clicking the mouse as well as realistic symbology and tab displays on the computer monitor. In addition, the navigation is similar to what game developers use. When the cursor approaches an edge of the screen, it turns into a "hand" pointed in that same direction, which when clicked with the mouse, enables the user to navigate to the next graphic. For example, when the Soldier is looking at Panel A55 (the console panel) and moves the cursor to the bottom of the screen and clicks with the mouse, the Soldier sees the computer keyboard that is physically located below the console panel.

Scoring. The simulation includes several scores that, taken together, evaluate the Soldier's performance. There are different levels of scores – the most important of which is whether the Soldier passes by correctly identifying the cause of the fault or exits early. In

general, the program records the number of steps the Soldier takes to solve the problem and any errors. Thus, low scores reflect effective performance and generally a higher level of expertise while high scores reflect errors or inefficient actions. An example of an error would be opening or operating an incorrect panel or switch.

Operational Assessment

Ideally, an operational test for the 14E MOS would include additional elements not built into the prototype. For example, an operational test would include enhanced multiple-choice questions to test declarative knowledge. There would be additional tasks we would simulate to more fully assess the MOS requirements. Consideration should also be given to incorporating time pressure with distractions and interruptions into the Soldiers' assessments.

The multiple-choice questions would include drag and drop and matching, some enhanced with audio/video clips. The simulation component could include self-contained scenarios with embedded multiple-choice questions, and/or one or more smaller simulations incorporating a common and multiple related scenarios. The simulation elements would enable us to measure activity that might otherwise damage expensive equipment as well as also likely eliminate the need for hands-on assessments.

When developing enhanced multiple-choice questions and/or simulations, the SMEs emphasized the importance of ensuring a high level of realism, which can relate to graphics and navigation discussed above, but also to presenting the test in the context of an operational mission. For example, the SMEs recommended that an assessment follow the same order as the Soldiers would normally experience in their qualification Table 8 and Table 12 exercises. This would start with a Patriot overview to include system capabilities and cable lengths, then reconnaissance, selection, and occupation of position, emplacement after a march order, initialization, operator actions, crew changeover, preventative maintenance checks and services (PMCS), baseline tactical weapons control computer unit diagnostics (TWUD), operator actions, and finally, march order.

Major Issues for an Operational Assessment

One issue related to developing an operational assessment for the 14E MOS relates to the existence of two subspecialties, operator and maintainer, and the strategic decision to assess MOS KSAs across the two distinct occupations rather than just within the separate specialties. The SMEs said that the most effective 14E Soldiers are operators *and* maintainers because when deployed and focused on mission, they need to fill in where/when needed and have a systems perspective. Further, when they become platoon sergeant they will have both sub-specialties reporting to them so to be effective leaders they must know the duties and requirements of all of their subordinates.

A continuing issue related to stand-alone or multiple-choice simulations is the importance of balancing realism with technical requirements and affordability. The SMEs want to see a high level of realism, but as we learned, that does not necessarily mean 3-dimensional – the images just need to look realistic and the switches, keys, and dials need to operate. In terms of technical requirements, it will be essential to work closely with the Army information technology experts both with respect to hardware and software requirements, but also for human technical support in

troubleshooting issues related to loading the software onto the personal computer. The challenge, of course, is to define the level of realism and technology that meets the budget constraints. Our initial stab at this balance was to start with the Army's current technical requirements and to develop 2-dimensional images that provide appropriate functionality and maneuverability. While the initial simulation development costs are high, the more uses that can be developed for the images and content, the lower the overall costs. This is a key consideration in evaluating the cost-effectiveness of this measurement method. For example, the console in the ECS is where the operator typically spends most of his/her time, so once that equipment is functional, all that needs to change is to provide the supporting software responses to operation of switch indicators. Similarly, the maintainer spends most of his/her time at the RS so once those images are developed, they can be re-used and the only additional development cost is whatever new panels need to be operational.

Another significant challenge with respect to the 14E MOS is the highly technical nature of the equipment. We were fortunate to have continuing access to seven SMEs who provided ongoing support. When developing assessment items, and particularly when developing a computer simulation, ongoing support from the same SMEs will be essential to ensuring that the relevant important details are correctly portrayed.

A related matter is the need to upgrade equipment and software. There is no regular schedule for upgrades, and in fact the equipment upgrades occur as new equipment becomes available. For example, presently the launchers are being upgraded for new missiles, so while some Soldiers are training on the old equipment others are operating the new configuration three-launcher stations. Similarly, operating procedures are upgraded, and there is the potential that the MOS will evolve over time. Therefore, it will be important to review the assessment at least annually for equipment/manuals upgrades as well as changes to the MOS.

Finally, there are several issues generally related to test scoring. Initial SME support of the project was based, in part, on the need to "filter out" the lowest MOS qualified candidates. Since the goal is to develop an assessment program that selects the best candidates for promotion, the SMEs suggested tailoring the difficulty to range from easier to harder. This is not practical for the multiple-choice items, as these will be developed for a particular skill level. However, for the simulation portion, one approach is to explore differentiating Soldier high ability levels by weighting competencies for mission complexity (Bennett, Schreiber, & Andrews, 2002). In addition, attention will have to be given to how to combine the scoring for the computer simulations with traditional multiple-choice items. Whitnah (2004) advocates giving greater weight to the simulations because they take longer to complete, require greater knowledge and skill, typically require combinations of multiple tasks, and actually test performance rather than knowledge or recall. Last, since the simulation resides on a personal computer and not on a remote server, procedures will have to be developed to accumulate scores in a central data repository for analysis, storage, and dissemination.

19K Armor Crewman

The 19K MOS was included in the PerformM21 research, in part because we knew that hands-on testing would be the preferred measurement method by the 19K NCOs and

we hoped to leverage existing training simulators and software options to develop more practical testing options. We found that the 19K SMEs were every bit as insistent on hands-on testing as we might have expected and that use of existing simulators (as previously discussed) was not a realistic option, at least at this time.

SME Reactions to Testing

Although most 19K Soldiers interviewed agreed that there was a need for a 19K MOS assessment tied to promotions, they were concerned that 19K offered some unique challenges that would make creating a fair test difficult. Specifically:

- The 19K SMEs felt strongly that any 19K MOS assessments must include performance tests as well as knowledge tests. SMEs expressed concern about “book smart” Soldiers being the ones promoted. They felt that all performance-based tasks (which are predominant in the 19K MOS) should only be tested hands-on.
- Many 19K Soldiers do not work regularly with tanks. For example, Soldiers in recruiting and instructor positions do not have much opportunity to serve as tank crewmembers. Also, many 19K Soldiers deployed in Iraq and Afghanistan are actually performing tasks outside their MOS, particularly infantry-oriented tasks (e.g., clearing buildings, providing dismounted security). The SMEs feared that recruiters, instructors, and deployed Soldiers would be penalized because their tanker skills would become rusty while serving in their current duty assignments.
- The SMEs expressed concern about how to handle the variety of 19K equipment in a test program. There was strong reaction that a Soldier be tested only on the tank and weapons system on which he has been trained. They suggested a modular test where the commander could choose the tank/weapon system that fit his unit’s Soldiers, and only questions or performance on that particular tank/weapon system would be included.
- The SMEs had reservations about providing feedback on test performance to units. There was concern about the enlisted test program being used as a performance indicator on commander’s and officer evaluations.

The Prototype Assessment

The prototype assessment includes (a) a sampling of the Select21 19K questions, (b) a simulation-based test of using the M1A2 SEP DID to check the engine health, (c) a simulation-based test of completing a function check on the .50 caliber machine gun, and (d) a series of animated multiple-choice items on the .50 caliber cold gun procedures.

Operational Assessment

An operational assessment for 19K would include (a) enhanced multiple-choice items that contain graphics, audio files, and animations to test knowledge-based skills; (b) simulation-based assessment items on the driver’s display and on setting up voice communication systems;

and (c) hands-on items to test performance on basic 19K performance skills like those found in the TCGST. However, it should be noted that SMEs were not enthusiastic about incorporating selected TCGST station tasks into the 19K MOS assessment. They were concerned that the local unit's control over how the TCGST is administered would be taken away if its tasks were incorporated into the 19K MOS assessment. The TCGST is currently perceived as an effective tool that might be tampered with because of the 19K MOS assessment development.

Major Issues for an Operational Assessment

Based on discussions with the 19K SMEs, following is a summary of major issues associated with technical promotion testing in this MOS.

- Given 19K equipment differences, it will be difficult to design tests that are suitable for everyone in the MOS. This might be done with (a) a universal test that is not tank-specific, (b) a test that can assess 19K skills on a common tank platform that every Soldier should know (perhaps the M1A1), or (c) tracked tests that are tailored to specific tank and weapon systems. If neither a universal test nor a common tank test can be produced, then creating tracked tests for three to four different types of tanks will be challenging.
- It is Army policy that Soldiers who are in non-MOS specific long-term assignments (such as recruiting, drill sergeants, or instructors) still are required to maintain MOS proficiency skills. The problem of non-MOS utilization during short-term deployments is also one that many MOS will face. Consideration should be given to a testing policy that avoids testing Soldiers during, and probably immediately after, deployments, allowing time for reacquisition of MOS skills.
- The 19K SMEs clearly favored a hands-on approach to testing their MOS. Implementation of a large-scale hands-on test presents challenges in test administration and comes with a predictably higher cost. Sustaining commitment to supporting operational hands-on testing in light of these difficulties can be a problem in all MOS for which hands-on testing might otherwise be appropriate.
- Although a hands-on approach is favored in the 19K MOS, the recommended approach is that a knowledge test be the basis for all MOS testing, to be supplemented by other forms of tests as appropriate. However, there is a legitimate concern on how hands-on scores will be weighted against knowledge scores. This is a concern in any MOS that uses more than one form of assessment, but it is magnified in the hands-on test because of the perception that hands-on tests are a more suitable measure.
- Army hands-on tests are usually scored pass/fail. Test scores used for promotion decisions, however, should be based on a point system. At least with 19K, there is expressed sentiment that Soldiers must be able to perform some tasks on pass/fail criteria in order to be promoted. This not only complicates the scoring and computation of promotion points but could cause other problems such as in retest policy. However, if there are no mandatory tasks a Soldier must be able to complete to be promoted in 19K, this may be difficult to reconcile given the requirement for each Soldier to pass all stations in the TCGST. Given that the 19K SMEs encouraging the development of an

MOS assessment saw it as a tool to help “weed out” Soldiers who cannot perform, letting Soldiers be promoted who cannot pass all the hands-on tasks appears to go against the SMEs’ reasons for advocating the assessment at all.

31B Military Police

The 31B MOS was selected for this project for two main reasons. First, 31B Soldiers are playing an increasingly critical peacekeeping/combat support role in current deployment activities (e.g., in Iraq). Second, the increased prominence of this role has led to perceptions that this MOS comprises two rather disparate missions—garrison law enforcement and peacekeeping/combat support—that may require very different KSAs. Indeed, many of the SMEs with whom we met are concerned that when garrison MPs are deployed in a combat support role, they often lack the tactical skills and experience to succeed in a combat environment.

SME Reactions to Testing

The SMEs with whom we met had a wide range of reactions to MOS-specific competence testing. For instance, many of the more experienced SMEs dismissed such testing as impractical and bound to face the same fate as the SQT. Other SMEs, in contrast, believe that the institution of a more objective competency testing system is critical for identifying effective leaders in this MOS. Despite the variation in opinions, SMEs generally indicated that there is a need for some form of MOS competency testing. In addition, SMEs agreed that to be successful, an assessment program would need to incorporate the following elements:

- Testing must assess the KSAs critical for success in both law enforcement and peacekeeping/combat support missions.
- Testing must incorporate both written and simulated or hands-on assessments to ensure that promoted Soldiers are “well-rounded.”
- At the same time, the testing situation must be such that deployed MPs have every opportunity to prepare for and take the tests in a suitable environment.

The Prototype Assessment

We considered these and other SME recommendations when designing the prototype assessments for this MOS. Although we considered a variety of potential assessments (including the Form 3975 assessment described earlier), in the end we developed two prototypes. The first was an SJT, which was selected for several reasons. First, our proponent told us that leaders within this MOS are very interested in situational testing. Second, SMEs indicated that situational judgment is a critical, but often overlooked, aspect of performance in this occupation. Specifically, 31Bs encounter critical situations on a daily basis that are not explicitly covered in training. Finally, the results of a preliminary job analysis suggested that interpersonal skills are very important for this MOS, yet SMEs indicated that such skills are not normally considered in promotion decisions. Thus, we concluded that an SJT would be a viable method for measuring interpersonal competence (although as discussed earlier, we found that SMEs had difficulty developing SJT items that assess interpersonal skills, per se.).

The second assessment we selected for this MOS was the previously discussed EST 2000. As with the SJT, the EST was selected as a prototype for a variety of reasons. First, it is an assessment method (i.e., a hands-on test) that our SMEs strongly encouraged us to pursue. Second, the EST can be used to assess various 31B KSAs (e.g., communication skills, tactical skills, judgment and decision making) critical to both law enforcement and combat support missions. And third, this assessment makes use of an existing simulation, thereby requiring relatively minimal time and resources to develop.

Operational Assessment

Based on what we have learned, the ideal operational testing program for this MOS would comprise multiple assessments. Minimally, it would include a multiple-choice test and an SJT. The multiple-choice test would assess the critical knowledge areas of the 31B MOS and the SJT would assess judgment and decision-making in the critical situations 31B Soldiers encounter. The SJT could focus on judgment in the core MP functions (e.g., law and order, area security) and/or be designed to assess specific KSAs (e.g., leadership and interpersonal skills) not explicitly measured by other elements of the competency assessment system.

In addition to these two assessments, we recommend that the 31B proponent consider some type of hands-on simulation for their assessment system, particularly for measuring tactical competence. As discussed, the current EST 2000 simulator is not ideal for use in a large-scale competency assessment system. Nonetheless, we received very positive reactions from SMEs about incorporating something like the EST 2000 into an assessment program, and we believe that the EST holds promise for serving this purpose.

Major Issues for an Operational Assessment

Of course, implementing an assessment system such as this would be challenging. There are two critical issues to be addressed prior to implementation in the 31B MOS.

Perhaps the main issue is how to weight the relative importance of law enforcement and combat support competence. Although there is not consensus that these two missions require fundamentally different KSAs, results of our initial job analysis provide some support for this belief. For example, interpersonal skills were rated as most critical for 31Bs in a garrison environment, whereas vigilance and psychomotor skills were rated as most important in peacekeeping/combat support situations. On the other hand, skills such as judgment and decision-making were rated as equally critical for both missions. In any case, the resolution of this issue would likely impact the focus of an assessment system.

Regardless of their relative importance, tactical skills are clearly essential to success as an MP and therefore must be evaluated in a competency program. However, available methods for assessing tactical skills are somewhat limited. For instance, we found it very difficult to develop SJT items that include the myriad of details needed to make an effective decision in a tactical situation. A simulator such as the EST 2000 might be a viable alternative for measuring tactical skills, although developing a secure set of combat-related scenarios will be time and resource intensive.

Despite challenges, the majority of SMEs with whom we have met concur that it is time to seriously consider the development of a more objective and fair system for selecting future leaders in this MOS.

63B Wheeled Vehicle Mechanic

The 63B MOS was selected in part because it is in a state of transition. MOS consolidation is prevalent and forecasted to continue in the foreseeable future. Identifying the problems and challenges encountered in testing an MOS in a state of instability was considered an important research objective. The new 63B is a result of combining three existing wheel vehicle mechanics jobs. This merging and combining has had the effect of de-specializing the 63B Soldier's job that has led to the adaptation of principles-based training for some of the same reasons we advocated principles-based testing. Dealing with an MOS going through such a transition resulted in some research problems. We were initially unable to obtain knowledgeable SMEs to participate in our work because they did not exist or were occupied with implementing the transition of training programs of instruction. This resulted in a delay in getting started with prototype test development for this MOS.

SME Reactions to Testing

The Army SMEs were receptive to the idea of an MOS technical testing program. More specifically, they liked the idea of a competency-based test. They acknowledged problems with the SQT, but if those could be overcome, they felt there was value in pursuing a new testing program because they perceive that there are too many ill-prepared Soldiers being promoted to E5.

The Prototype Assessment

The prototype assessment includes approximately 90 job knowledge items. The majority are competency- or systems-based items covering the topics of Electrical Systems and Engines. The items make liberal use of graphics and the amount of text examinees need to read is limited.

Operational Assessment

Hands-on testing arguably may be one of the best methods of testing these Soldiers. However, as noted by R.C. Campbell (1994), the need for scorers and equipment were among the problems with the old SQT. The 63B SMEs echoed this concern. A high need for equipment and/or scorers is likely too much for the Army to support in the current environment of constant troop deployments.

For an operational test for the 63B MOS, we recommend a combination of hands-on testing (if prior issues can be overcome) and a competency-based job knowledge test. This combination allows the capture of the combination of declarative and procedural knowledge necessary for successful performance as a 63B.

An Issue for an Operational Assessment: Civilian Certifications

There are related civilian certifications for this MOS. Soldiers are awarded 10 promotion points for each certification earned, not to exceed 50 points. The Medium/Heavy Truck Technician certification offered by the National Institute for Automotive Service Excellence (ASE) is the most common civilian certification endorsed by this MOS. Other advanced civilian certifications for which 63B Soldiers may earn promotion points are offered by the American Society for Quality and the Board of Certified Safety Professionals.

According to the ASE website, approximately 420,000 automotive professionals hold ASE certifications. There are more than 40 different exams grouped into the following specialties:

- Automobile
- Medium/heavy truck
- Truck equipment
- School bus
- Collision repair
- Engine machinist
- Alternative fuels technician
- Parts specialist
- Auto service consultant
- Collision damage estimator

A person with 2 years of relevant work could obtain certification by passing at least one test. In order to keep the certification current, the test must be repeated every 5 years. The majority of these exams are paper and pencil tests, but a computer-based testing (CBT) pilot is being conducted.

The 63B MOS awards promotion points for Soldiers completing portions of the ASE civilian certification. The ASE certification was neither developed by nor is it administered by the Army, which not only means cost-savings, but also reduces the administration burden. ASE certification is available now, and it covers the same topics included in the blueprint we developed. However, there are some conditions that must be considered in adoption of civilian certification. For instance, Soldiers are currently allowed to take whichever module(s) they like, thereby allowing them to ignore areas that may be critical, but difficult, for them. Similarly, Soldiers take one module at a time, demonstrating competency in one area at a time. In contrast, an Army-developed test would likely require demonstration of competency in a broader range of areas simultaneously. Finally, one either passes or fails a certification test; there are no point scores. For an operational test, we recommend giving Soldiers point scores rather than using a specific cut-off.

Civilian certification and competency assessment should not be viewed as an "either/or" situation. It is possible to have both, which may require some integration, especially with respect to how promotion points are awarded. Other areas of integration may include allowing Soldiers to exempt small portions or modules of an Army test if they have a current ASE certification in a

related area. An Army test may encourage more Soldiers to take advantage of the certification program. As mentioned above, Soldiers will have to demonstrate competency in a broader range of areas simultaneously, which may lead them to pursue ASE certification in several areas. Finally, it is quite possible that the test preparation/self-development tools offered by ASE would be useful to help Soldiers prepare for an Army-specific assessment.

68W Health Care Specialist

Among the reasons for selecting the 68W MOS was the existence of relevant civilian certifications and the potential for principles-based testing. This MOS was also particularly useful for exploring adaptation of existing Army tests and training simulators for promotion testing purposes.

SME Reactions to Testing

The 68W SMEs that participated in the workshops were originally rather negative about the idea of testing. Their reactions included the following:

- “Book smart” does not necessarily make a good 68W.
- 68W Soldiers are tested enough already with the semi-annual combat medic skills verification test (SACMS-VT) and the requirement to keep their emergency medical technician – basic certification (EMT-B) current.
- There is a huge difference between the “straight medic” and one with an M6 Additional Skill Identifier (licensed practical nurse [LPN]).

The concern about “book smarts” was the most frequently mentioned complaint about testing. Soldiers’ comments were replete with examples of 68Ws who tested very well in AIT, only to fail miserably when faced with an actual patient in a real-life situation.

Soldiers in the 68W MOS are required to keep their EMT-B certification current, which means they must reregister every 2 years. EMT-B is run by the National Registry of Emergency Medical Technicians (NREMT), a civilian organization most closely associated with paramedics or emergency medical service providers in the civilian sector. Promotion points are given for earning certification.

The Army also requires that Soldiers take and pass the SACMS-VT, which is supposed to be administered twice a year. It presents Soldiers with simulated common medical emergencies (e.g., heart attack, gunshot wound), and requires them to provide treatment. This is essentially a hands-on work sample test, using either sophisticated mannequins or other Soldiers to simulate injuries/illness. Typically, two administrators provide ratings for each Soldier in each scenario. Soldiers who are current with their SACSM-VT fulfill the skills maintenance requirements for the EMT-B (Department of the Army, 2002). Depending on the train-up they engaged in for the SACMS-VT, they may or may not need continuing education units (CEUs). There are some locations that do not have SACMS-VT, or who do not administer it semi-annually as required.

For instance, it is currently available in all Medical Command (hospital) units, but will not be available in divisional (troop) units until 2007.

The SACMS-VT consists of seven training “tables” and the overall test, which covers skills from these tables (see Table 27). In order to pass the test, a Soldier must earn 70% of the points on the skills covered in Tables I – VI, 100% of the skills covered in Table VII, and not miss any of the critical skills. An example of a critical skill for Trauma Assessment is *Determines number of casualties* (Department of the Army, 2002).

Table 27. Design of the SACMS-VT

	Table	Covered Skills
Training	I	Trauma Assessment and Management
	II	Immobilization of Bone and Joint Injuries/Extraction
	III	Medical Assessment and Management
	IV	Basic and Advanced Airway Skills
	V	CPR Management
	VI	NBC Medical Skills
	VII	Evacuation
Testing	VIII	Hands-on Testing of Tables I - VII

The 68W is the designation for the newly transitioned 91W MOS. The 91W MOS was a consolidation of 91B (Medical Specialist) and 91C (Practical Nurse). 91C Soldiers who transitioned over have an M6 Additional Skill Identifier (ASI). This consolidation has been quite a challenge for these two diverse MOS, one whose background is a hospital (91C) and the other whose background is in the field (91B). Many of our SMEs felt these differences were huge and would make it nearly impossible to find enough overlap to justify a common test. However, their SACMS-VT is able to accomplish this, even if it is not currently being used as required by TC 8-8000, Semi-Annual Combat Medic Skills Validation Test (SACMS-VT).

Anecdotal evidence suggests that the proper train-up (i.e., completing the seven training tables) is not being conducted. Soldiers in the 68W MOS may work in a variety of settings/units (e.g., hospital, aid station, forward support battalion, or battalion medical evacuation platoon) that differ in the amount of practice or actual performance of 68W tasks they provide. For instance, a 68W in a hospital will necessarily have more practice than one assigned to an infantry brigade who spends most of his time in the motor pool. If the proper train-up for the SACMS-VT is not conducted, then many Soldiers will need a few practice runs. Therefore, often Soldiers are run through the test portion (Table VIII) until they achieve 100%, which is then recorded as their official score.

Prototype Assessment

The prototype assessment developed for this MOS consists of two parts: (a) an SJT and (b) a job knowledge test. Once finalized, the SJT will likely consist of 20 to 25 items covering emergency medical care/triage and interpersonal communication. The graphically-enhanced job knowledge test has a competency-based set of questions and a task-based set of questions.

Operational Assessment

For an operational testing program for the 68W MOS, there should be a knowledge- or competency-based portion plus a hands-on portion, which is supported by our 68W MOS proponent POC. In fact the POC is so serious about testing knowledge areas and competencies that he has instituted such testing for the 68W AIT instructors. Any instructor scoring less than 80% on the test he or she developed is reassigned.

The knowledge-based portion could be comprised of job knowledge questions and an SJT, similar to what was done for the prototype assessment. The hands-on portion could be based on a new “testing” SACMS-VT incorporating the SimMan® simulators (discussed below). We would suggest that a new “testing” SACMS-VT include the following:

- The proper train-up as required
- A focus on standardized presentations across locations
- Properly trained raters to serve as assessors
- Point scores as opposed to a pass/fail

Training Circular (TC) 8-8000 governs the administration of the SACMS-VT and is written for both the test-takers and commanders who are in charge of implementation. It is fairly detailed, providing sample training and testing scenarios as well as specifying which skills are to be tested. However, each commander is responsible for creating his own scenarios. TC 8-8000 specifies that for each station (scenario) there should be two evaluators – one to provide the rating, and one to participate in the scenario. But, the SMEs suggested to us that is not always the case. It would be desirable for there to be two evaluators or raters per scenario, each providing a rating. We realize this contributes to the resources issue for hands-on testing. However, there are a couple of reasons why two raters are better than one. First, these assessments are conducted in “real time” and there is often a lot going on. It is easy for one rater to miss something important. Second, it allows for more sophisticated analyses for evaluating the testing program and rater performance.

For SACMS-VT, any qualified 68W (i.e., one who has completed the transition) is able to serve as an evaluator. There are instructions for evaluators (e.g., do not coach the examinees), but from what we were told, rarely are any kind of rater training sessions conducted. We would recommend that rater training be performed prior to each SACMS-VT administration whether or not it is included as a component of competency assessment.

Finally, as mentioned before, passing means earning 70% of the points on the skills covered in Tables I – VI, 100% of the skills covered in Table VII, and not missing any of the critical skills. We would suggest moving from a pass/fail format to one in which Soldiers earn points commensurate with their performance, which would then be applied to their promotion point worksheets.

There is interest in reducing the requirement for administering the SACMS-VT from semi-annually to annually. This could free up some resources to be devoted to making this approximate a more full-fledged testing (as opposed to training) situation (i.e., conducting the

proper train-up so multiple runs through the test portion are not required). This would mean that most units would have had several years with this testing format so a long rollout would not likely be needed for a new or modified "testing" SACMS-VT.

Major Issues for an Operational Assessment

The biggest issue associated with instituting an operational test program in this MOS concerns the existing testing requirements (i.e., the SACMS-VT and EMT-B), which is a legitimate complaint. Many of the concerns voiced about testing revolved around the fact that they are already tested frequently. As mentioned, 68Ws are required to keep their EMT-B certification current. This certification expires every 2 years, and the SACMS-VT is designed to help in this process. Some of the SMEs we spoke with feel the EMT-B certification should be the only testing required in this MOS. However, re-registration is not necessarily testing. It consists of the following:

- 24 hours of didactic work covering areas such as airway, patient assessment, or trauma which may be satisfied with an approved EMT-B refresher training course or equivalent continuing education courses
- Continuous, current CPR certification
- 48 hours of continuing education courses from subject matter covered in any EMT-B (or higher) course
- EMT-B skills maintenance (National Registry of Emergency Medical Technicians, n.d.)

We were told by SMEs that some training (not *testing*) conducted at the squad level may count towards the continuing education requirements if documented properly. Also, the SACMS-VT specifically covers the 48 hours of continuing education courses and verification of skills maintenance of the EMT-B re-registration requirements. If the SACMS-VT were administered as suggested above, it could fulfill the requirements for a procedural knowledge test while still helping to alleviate the re-registration requirements for the EMT-B certification. In other words, it would have little impact on the current testing requirements. Further, the knowledge-based portion would not necessarily have to be lengthy.

Aside from using an updated "testing" SACMS-VT as part of a testing program, there is the possibility of using the computer-based MicroSim® program.

Summary

A test of Soldier competency that does not include job-specific measures is not a complete test. Technical competency is the essence of most Soldier performance requirements and must be considered as a basic premise of evaluations and personnel actions including promotion decisions. Yet, this is the most challenging and will be the most difficult part of Soldier testing to implement. There is great diversity in Army jobs and the issues faced are often specific to a particular job making it difficult to apply universal solutions or approaches.

The approach in PerformM21 was to examine five representative MOS to see what was feasible, what problems existed, and what resources could be applied. The aim was not to perfect a particular job assessment but to contribute to the body of knowledge and experience about Army testing in job-specific applications. It must be noted that the PerformM21 work was done with relatively limited resources and does not replicate the climate that would exist within a proponent if an operational MOS test development priority were established. In our approach, we operationalized the test under current conditions and tried to project what would be possible in the future. The realities, in a stood-up system, may be far different, both in challenges and in opportunities. But the experience has identified some important developmental tenets for MOS specific testing:

- MOS are different. In the old SQT testing, all MOS were treated essentially the same in how they were tested. Recognizing that this will not be a requirement of a new test system is a fundamental. Proponents should be encouraged to seek innovative and unique testing applications and solutions to match the conditions of the job. This does not mean there would be no rules, boundaries, or uniform structures. An overall testing authority must still ensure that sound testing principles and testing efficiencies are adhered to, but there is still room for individualized solutions and test methods.
- A basic foundation for all MOS should be a job knowledge test. Some aspect of job knowledge is a requirement in every job and job knowledge tests are comparatively easy to develop and administer. Job knowledge tests may not be sufficient or the sole testing method for all MOS, but they should be the start point.
- The interest generated on the part of the five proponents was significant. Proponents want a voice and an active role in advancing testing investigation and in developing test policy. This does not mean that there is universal enthusiasm for testing. Many doubts and concerns still apply. But there is a high interest in participation.
- The application of different test methods for different performance requirements is a viable concept and developmental work should continue. These methods include simulations, work products, utilization of existing tests and certifications, hands-on tests, and the conversion of training aids and devices to test applications. Computer-based testing can offer capabilities that erase the distinction between knowledge-based and performance testing.
- Assignment-specific equipment, tasks, and missions is an almost uniform issue in all MOS, albeit with individual variations and applications. In order to avoid a cumbersome and expensive test system, this will require innovative approaches to test design and content. The detailed task-based training system, with which the Army works and is most familiar, is probably not the approach that will be needed in a test system. This will require a shift in thinking by proponents and SMEs.
- Many of the problems, restrictions, objections, and difficulties we encountered are issues that could be resolved over time or with sufficient resources. MOS test development should be incremental, starting with the more feasible testing

approaches and developing more sophisticated and perhaps complex test methods and applications over time.

- Job analysis, specifically tailored to address test issues, remains a paramount requirement. SME input in test content and application is invaluable, but risks parochialism, bias, and lacks systematic consideration of all factors. The structuring of job analysis to support test requirements needs more attention and formalization.
- Many individuals' perceptions of a test program are based on anecdotes from the SQT era, experiences with training tests, or assumptions about what a test program would include or how it would operate. They are important, because perceptions guide outlooks and attitudes, but it is also important that they not overly influence thinking or progress. No decisions have been made and no policy is in place, but this reinforces the requirement for planning and widespread participation in an Army test program.

CHAPTER 6: SUMMARY AND NEXT STEPS

Deirdre J. Knapp and Roy C. Campbell

PerformM21 Accomplishments to Date

At the end of Phase II of the PerformM21 research program, we have a reasonable idea of what a new Army competency assessment program might look like, including test design, development, delivery, and maintenance, as well as considerations related to policy development and management of the program. We have developed a prototype core assessment targeted to Specialists/Corporals (E4s) eligible for promotion to Sergeant (E5). We have administered this core exam to a sample of almost 600 Soldiers using existing Army distance training facilities as test centers. We have taken the prototyping part of the research a step further by designing and developing job-specific assessments for several MOS. This is allowing us to tryout different job analysis strategies and different assessment methods (e.g., computer simulations).

Thus far, we have learned nothing to indicate a new assessment program is not an idea worth continued serious consideration. Concerns voiced by SMEs who have participated in our test development efforts should be taken seriously, but are not insurmountable in a carefully designed assessment program. It is also clear, however, that Army leadership needs to understand the scope of commitment required to support such a program before making the final decision to do so. This includes an organizational long-term commitment of substantial resources, communicating with stakeholders to address their concerns and obtain their cooperation, and commitment to quality so that Soldiers are subjected to tests that are both valid and fair. That said, if the Army moves ahead with a new assessment program, it will have successfully responded to key recommendations from the ATLDP participants and fulfilled ideas fostered by many senior NCOs including the Sergeant Major of the Army. Moreover, the positive effects of the program will be evident in improved force readiness. This is, after all, the ultimate goal of the program and what most makes it worth pursuing.

PerformM21 Phase III

In Phase III of the PerformM21 project, researchers will administer the prototype MOS assessments to samples of Specialists/Corporals. We will use their test data to evaluate and improve the test questions as well as gain new insights into test methods and options. We will also ask participating Soldiers for feedback on the entire assessment process (e.g., signing up to take the test, usefulness of the test preparation guide) and use this to improve elements of the recommended operational program.

We will continue to work with the ATPAT and other resources in Phase III to develop options for how an operational competency assessment program would be designed and implemented. We will endeavor to provide enough information about program requirements to help determine the costs and benefits of adopting the program (in all or in part). The ultimate goal of the project is to help the Army make fully informed decisions about a major new initiative.

An Army-Wide Core Competence Assessment

As described in Chapter 3, PerformM21 has produced a bank of test items suitable for an Army-wide core assessment, as well as a test preparation guide, test delivery protocols, and a Soldier feedback report. As part of a separate effort, a prototype self-assessment program geared to the Army-wide core competency test (and the LeadEx) was developed.

The Army-wide core pilot test was well received by the Soldier participants. Analysis of performance data reveals that there were differences in test performance across racial/ethnic groups. Every effort should be made to mitigate such differences in an operational program (e.g., through careful test development and adequate examinee preparation), but historical examples indicate that all differences are unlikely to disappear. There were a number of technological glitches in the pilot test that would be unacceptable in an operational program, suggesting that the promise of technology cannot be realized without considerable effort to make it all work as envisioned. In Phase III, we will examine in more detail the cost drivers and expected benefits of a core competence assessment program. A separate project has just begun to develop and pilot test additional core test items based on lessons learned in recent deployment activity in the Middle East.

There are a number of steps that would be required before the Army-wide core test could be implemented. The following list is illustrative rather than exhaustive. It also assumes the Army would pursue implementation of both the core examination and the LeadEx for testing done in support of semi-centralized promotion decisions.

- Develop a core exam blueprint based on a thorough job analysis that reflects policy decisions about test content (e.g., inclusion of skill levels higher than the target examinee pay grades and content that would not be captured by a typical job analysis, such as history and values).
- Create a much larger bank of both core exam and LeadEx items that allows for creation of multiple equivalent forms.
- Devise details of a process for creating equivalent core exam and LeadEx forms within and across administration periods.
- Develop scoring protocols that include checks on test quality and possible test compromise.

Technical (MOS-Specific) Competence Assessment

By the end of Phase III, PerformM21 will have produced a good deal of information that the Army can use to determine a path to pursue with technical testing, at least for some MOS. Even without results of the cost-benefit analysis, it is evident that MOS testing would greatly increase the management complexity and costs of having an assessment program. It is also evident that a one-size-fits-all approach to technical testing is neither feasible nor desirable. Indeed, we recommend starting with a small number of MOS that have the greatest need or interest in testing and develop programs that reflect their requirements. But it is with technical

testing where the greatest proponent interest is and where ultimately the greatest performance payoff may be.

Closing Comments

An Army test program cannot be viewed as a separate or isolated program. If the Army elects to proceed with testing, it should be because it is a part of an integrated Soldier and NCO development plan that includes training, self-development, experience, utilization, mentoring, and promotion of the best qualified and proven individuals. We believe that testing has a role in such a plan and that it can be successfully integrated with other facets of career development, enhancing rather than restricting utilization of the Army's human capital. Assessment will not be a universal answer to all the Army's problems; it must be applied judiciously and with well-understood and sensible expectations of what it is meant to do. But it can be a valuable tool if used correctly and applied to specific goals.

Initiation of an Army test is a major undertaking, without a doubt. But so too are many of the other initiatives that are currently being considered for the 21st century Soldier. The timing is right and so is the opportunity. Unlike the sister services which are trying to change legacy personnel testing systems, the Army has the opportunity to start with a clean slate, avoiding problems from the past and taking advantage of new technologies and new testing concepts. An Army test system that is well planned, stable, properly resourced, fairly administered, and integrated with other requirements can be a model program. Our goal has been to provide the foundation for such a program.

REFERENCES

- Aaby, A. (2001) *Software maintenance* Open Content
<http://cs.wwc.edu/~aabyan/435/Maintenance.html>
- Baisden, A., & Rymsza, B., (2004) *New directions in Navy assessment: Developing a multimedia prototype*. Paper presented to the 48th Annual Conference of the International Military Testing Association. Brussels, Belgium.
- Bennett, W. Jr., Schreiber, B. T., & Andrews, D. H. (2002). Developing competency-based methods for near-real-time air combat problem solving assessment. *Computers in Human Behavior*, 18, 773-782.
- Budescu, D.V. (1988). On the feasibility of multiple matching tests – Variations on the theme by Gulliksen. *Applied Psychological Measurement*, 12, 5-14.
- Campbell, J.P., & Knapp, D.J. (Eds.) (2001). *Exploring the limits in personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Campbell, R. C. (1994). *The Army Skill Qualification Test (SQT) Program: A Synopsis* (HumRRO Interim Report IR-PD-94-05). Alexandria, VA: Human Resources Research Organization.
- Campbell, R.C., Keenan, P.A., Moriarty, K.O., & Knapp, D.J., & Heffner, T.S. (2004). *Army enlisted personnel competency assessment program Phase I (Volume II): Demonstration Competency Assessment Program development report* (Technical Report 1152). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cullen, M. J., Sackett, P. R., & Lievens, F. (2004, April). *Threats to the Operational Use of Situational Judgment Tests In the College Admission Process*. Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Orlando, FL.
- Department of Defense. (2002). *Defense integrated military human resources system (Personnel and pay) DIMHRS (Pers/Pay)*. Report to Congress. Washington, DC.: Author.
- Department of the Army (2002). *Semi-annual combat medic skills validation test (SACMS-VT)* (Training Circular 8-800). Washington, DC: Author.
- Department of the Army. (2002). *The Army training and leader development panel report (NCO)*. Final Report. Fort Leavenworth, KS.; U.S. Army Combined Arms Center and Fort Leavenworth.
- Drewes, D.W. (2000). Beyond the Spearman-Brown: A structural approach to maximal reliability. *Psychological Methods*, 5, 214-227.

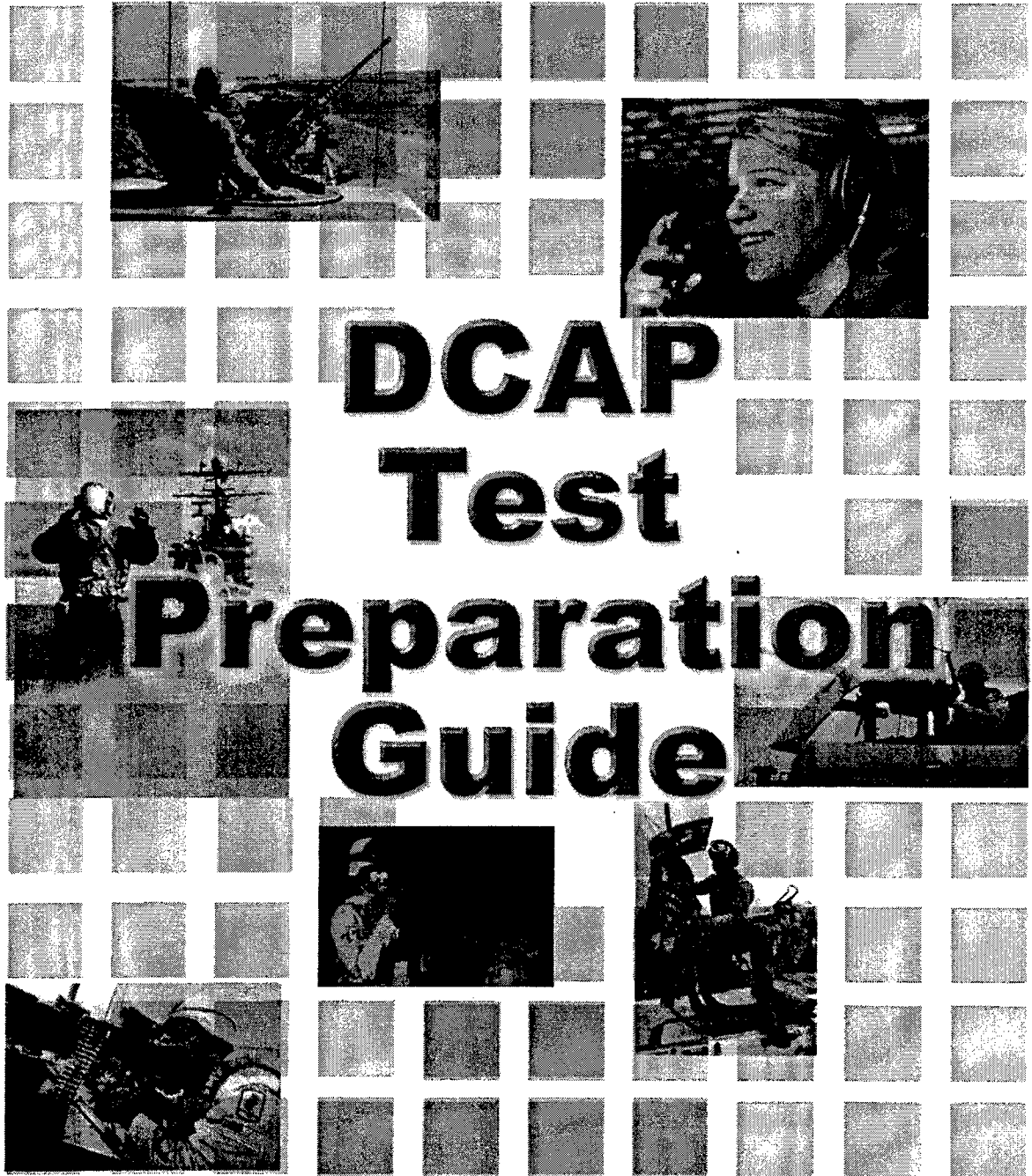
- Flanagan, J.C. (1954). The critical incident technique. *Psychological Bulletin*, 51, 327-358.
- Ford, L.A., Campbell, R.C., Campbell, J.P., Knapp, D.J., & Walker, C.B. (2000). *21st century Soldiers and noncommissioned officers: Critical predictors of performance* (Technical Report 1102). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Freudenrich, C. (2004) *How fiber optics work*. Verizon Learning Center. New York, NY.
- Keenan, P.A., & Campbell, R.C. (2005). *Development of a prototype self-assessment program in support of soldier competency assessment (FR-04-80)*. Alexandria, VA: Human Resources Research Organization.
- Knapp, D.J. (Ed.) (2003). *Select21 measure development progress report* (Interim Report 03-74). Alexandria, VA: Human Resources Research Organization.
- Knapp, D.J., & Campbell, R.C. (2002). *Selection for leadership: Transforming NCO promotion* (Special Report 52). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D.J., & Campbell, R.C. (2004). *Army enlisted personnel competency assessment program phase I (Volume I): Needs analysis* (Technical Report 1151). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D.J., Burnfield, J.L, Sager, C.E., Waugh, G.W., Campbell, J.P., Reeve, C.L., Campbell, R.C., White, L.A., & Heffner, T.S. (2002). *Development of predictor and criterion measures for the NCO21 research program* (Technical Report 1128). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D.J., Heffner, T.S., & Campbell, R.C. (2003). *Recommendations for an Army NCO semi-centralized promotion system for the 21st century* (Research Report 1807). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Knapp, D.J., McCloy, R.A., & Heffner, T.S. (Eds.) (2004). *Validation of measures designed to maximize 21st-century Army NCO performance* (Technical Report 1145). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology*, 87, 675-686.
- McDaniel, M.A., Moreson, F.P., Finnegan, E.B., Campion, M.A., & Braverman, E.P. (2001). Use of situational judgment tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology*, 86, 730-740.
- Military Personnel Management, HQDA Army G-1 (2004). *Army Training Requirements and Resources System (ATRRS) General Information*, <https://www.atrrs.army.mil/infor/atrrsinfo.asp>

- Motowidlo, S.J., Dunnette, M.D., & Carter, G.W. (1990). An alternative selection procedure: The low-fidelity simulation. *Journal of Applied Psychology*, 75, 640-647.
- National Guard Bureau (2004). *Distributive Training Technology Project (DTTP): DTTP Monthly Site Usage, October 2004 Usage, Recap and Review*, Washington, DC: Author.
- National Registry of Emergency Medical Technicians (n.d.). *Reregistration requirements for NREMT basic*. Columbus, OH: Author.
- Neidig, R.D. & Neidig, P.J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology*, 69, 182-186.
- Office of the Under Secretary of Defense, Personnel and Readiness (2003). Washington, DC. *Population Representation in the Military Services Fiscal Year 2002*
<http://www.humrro.org/poprep2002>.
- Peterson, N.G., Mumford, M.D., Borman, W.C., Jeanneret, P.R., & Fleishman, E.A. (Eds.) (1999). *An occupational information system for the 21st century: The development of O*NET*. Washington, DC: American Psychological Association.
- Program Management Office, Distance Learning System (2004) *Experience a training revolution*. Newport News, VA: Author.
- Rosenthal, D., Sager, C.E., & Knapp, D.J. (2005). *A strategy to produce realistic, cost-effective measures of job performance* (Study Note 2005-03). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Sackett, P.R., Schmitt, N., Ellingson, J.E., & Kabin, M.B. (2001). High-stakes testing in employment, credentialing, and higher education: Prospects in a post-affirmative-action world. *American Psychologist*, 56(4), 302-318.
- Schultz, K., Sapp, R., & Willers, L. (2003). *Electronic advancement exams – transitioning from paper-based to electronic format*. Paper presented to the 47th Annual Conference of the International Military Testing Association, Pensacola, FL.
- U.S. Army (2004) *TRADOC LMS Fielding Schedule*. Fort Eustis, VA: Author.
- U.S. General Accounting Office (1982). *The Army Needs to Modify its System for Measuring Individual Soldier Proficiency* (FPCD-82-28). Washington, DC: Author.
- U.S. General Accounting Office (2003). *Progress and Challenges for DoD's Advanced Distributed Learning Programs* (GAO-03-393). Washington, DC: Author.
- U.S. General Accounting Office. (2003). *Defense Management: Army Needs to Address Resource and Mission Requirements Affecting its Training and Doctrine Command* (GAO-03-214). Washington, DC: Author.

- Viswesvaran, C., & Ones, D.S. (1995). Theory testing: Combining psychometric meta-analysis and structured equations modeling, *Personnel Psychology*, 48, 865-885.
- Wainer, D. & Thissen, H. (2001). True score theory: The traditional method. In David Thissen (Ed.) and Howard Wainer, H. (Eds.). *Test Scoring*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Walcott, SFC (2004). *NCOES Briefing Slides*, Human Resources Command, Training Division, Specialized Training Management Branch.
- Whitnah, D. (2004). *A practical guide to simulation development*. TestOut Corporation.
Available Email: dwhitnah@testout.com.

APPENDIX A

DCAP Test Preparation Guide



■ Table of Contents

Introduction.....	A-3
General Information	A-3
What is PerformM21?.....	A-3
Why should I be interested?.....	A-4
How will my answers be used?.....	A-4
How to prepare.....	A-5
Testing day.....	A-5
What is Covered on the Assessment.....	A-6
Content Areas and Weights	A-6
Study References.....	A-7
The Army Noncommissioned Officer Guide.....	A-7
Army Leadership: Be, Know, Do	A-7
Sample Items	A-8
LeadEx	A-9
Test-Taking Strategies.....	A-10
Feedback.....	A-10
What kind of feedback will I receive?	A-10
When will I receive my feedback?.....	A-10
How will I receive my feedback?	A-11
Contact Us	A-11

■ List of Tables

Table 1. Study References	A-7
---------------------------------	-----

■ List of Figures

Figure 1. Sample Items	A-8
Figure 2. A Sample LeadEx Item	A-9

■ Introduction

You have been selected by your unit to participate in a research-only, pilot program for the Army's PerformM21 project. You have the opportunity to view the future of Army testing. Your participation involves completing an online prototype test at your local digital training facility (DTF) or digital training technology project (DTTP) location. Your time and effort are greatly appreciated. This preparation guide will explain the PerformM21 project further and help you prepare for your participation.

■ General Information

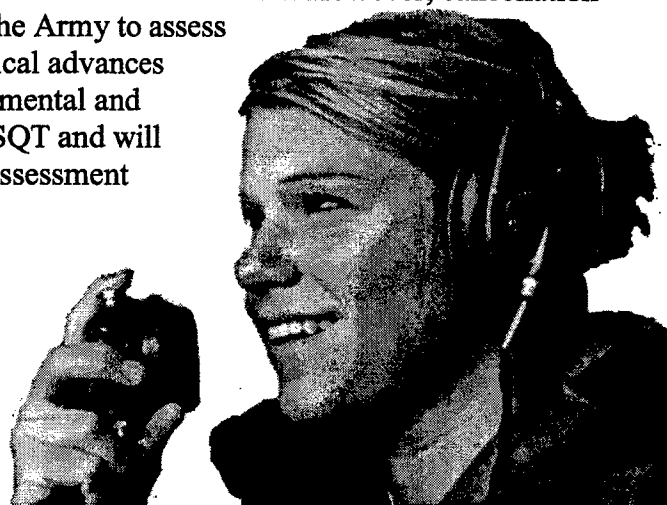
As you are aware, the Army is changing to meet the needs of the 21st century. Soldiers at all levels must possess the interpersonal, technical, and organizational skills to perform effectively in the Future Force. The Army needs an integrated Soldier assessment system to support these demands.

This type of assessment is most needed at the time of promotion into the non-commissioned officer (NCO) ranks. It is there technical skill merges with leadership and supervisory requirements, leading to distinct changes in the concept of Soldiering. A recent survey of 35,000 NCOs resulted in the following recommendation: "Develop and sustain a competency assessment program for evaluating Soldiers' technical and tactical proficiency in the military occupational specialty (MOS) and leadership skills for their rank."¹⁶ This recommendation prompted the Army to begin looking at instituting a Soldier competency assessment test as part of the requirement for promotion to the NCO ranks. The PerformM21 project is the result.

What is PerformM21?

In the early 1990s, the Army dropped its Skill Qualification Test (SQT) program due primarily to maintenance, development, and administration costs. However, cancellation of the SQT program made it difficult for the Army to assess job performance qualification. Technological advances have occurred that can reduce the developmental and administrative burdens encountered with SQT and will play a critical role in a new performance assessment system.

The purpose of PerformM21 is to determine if what has been learned since the days of the SQT can be put into practice. To do this, we have developed a knowledge test of subjects that all Soldiers, regardless of their job or



¹⁶ Department of the Army (2002). *The Army training and leader development panel report (NCO)*. Final Report.

MOS, are supposed to know. These subjects fall in the general areas of Common Tasks, Leadership, Training, and Army History, Customs, and Army Values. It is this test that you are being directed to participate in.

This pilot test is being administered to a select sample of Soldiers for experimental purposes only. All answers are confidential. Although your results on the test will be made available to you, these results or your answers will not be reported back to your unit, nor will they become part of your record.

Why should I be interested?



The Army is serious about reinstituting competency assessment as one of the tools in measuring fitness for promotion. This project has the support of the Army G1, TRADOC, and the Office of the SMA. You have an opportunity to play a role in this research and get a sneak peak into the future Army. Although a promotion assessment test will not be operational for several years, this assessment may be used with Soldiers you will one day supervise.

This test is also an opportunity for you to find out how you measure up today compared to other Soldiers like you – all unofficially of course. The pilot test will be administered to between 600 and 1000 E4 Soldiers who have about 2 years service, and who are from a mix of combat, combat support, and combat service support MOS. These Soldiers are being selected from the active Army, the US Army Reserve, and the Army National Guard. Part of your feedback will include how well you did relative to others who participated in this program. In other words, you can see “how you stack up” against other Soldiers in the sample in today’s Army.

Good data are needed to identify the best test items. These data will come from you and your fellow Soldiers selected for this pilot test. So, while this test is not reportable and does not really “count,” it IS very important that you prepare for and complete this assessment seriously. Collecting good data now ensures the operational assessment is well constructed and fair. As part of this, you are being given an opportunity to prepare yourself to take this test just as we think Soldiers in the future will prepare when the test becomes operational. That is the purpose of this Preparation Guide. Later sections present a list of references to review as well as the content areas covered on the assessment.

How will my answers be used?

As mentioned previously, this is an experimental administration, and your answers to these items are confidential. Only you will see your results. HumRRO (a contractor working with the Army Research Institute for the Behavioral and Social Sciences [ARI]) will combine the data and run various analyses to determine the quality of the items. Reports of these analyses will only include aggregated data, not individual data. Some participants will also be asked to provide verbal feedback about the test and the entire

process. This verbal feedback is also confidential and a vital part of the process to make this assessment a success.

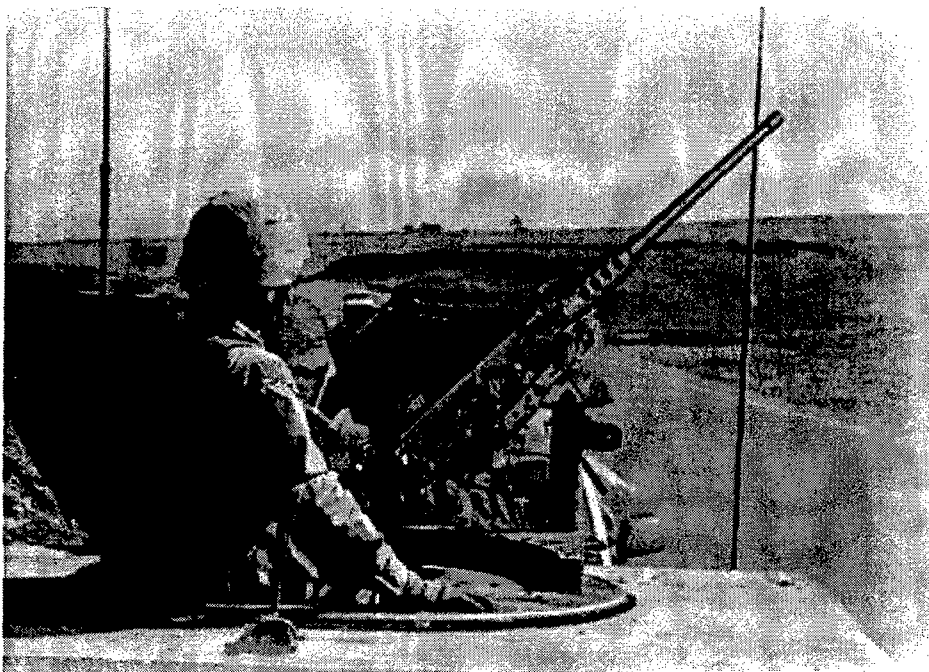
How to prepare

After reviewing the section, What is Covered on the Assessment, determine the areas you know well and those you do not. Focus on those you do not know well and establish a schedule to review the manuals or sections of manuals that address those areas. The Reference section of this guide contains the manuals from which these items were written. Simply click on the title to access the manual.

Do not wait until the night before the test day to begin your preparation. Rather, set aside some time every few days to review the material. If there are other Soldiers in your unit that are scheduled to participate in this test, you may study or prepare together if you desire. You may also ask your supervisor or unit NCOs for assistance in preparation by providing them the guidance contained in this Preparation Guide.

Testing day

The pilot test will take four hours to complete. On testing day, arrive at the DTF/DTTP 15 minutes prior to the start time. You will be required to present your identification. This test is completely computer based and administered. No food, drinks, cell phones, or pagers are allowed during the testing session. Nor will you be allowed to have any references or to take any notes during the test.



■ What is Covered on the Assessment

Content Areas and Weights

There are 5 content areas weighted as follows:

- Skill Level 1 Common Tasks – 46% of the assessment
- Skill Level 2 Common Tasks – 14% of the assessment
- Army History/Army Values – 15% of the assessment
- Leadership – 13% of the assessment
- Training – 12% of the assessment

Skill Level 1 Common Tasks (46%)

Skill Level 1 Common Task items are drawn from the *Soldier's Manual of Common Tasks Skill Level 1* (SMCT SL1). Tested subject areas include First Aid, Weapons (e.g., M16 and M9), Nuclear, Biological, and Chemical (NBC), Communicate, and Survive.

Skill Level 2 Common Tasks (14%)

Skill Level 2 Common Tasks are drawn from the *Soldier's Manual of Common Tasks Skill Levels, 2, 3, and 4* (SMCT SL2). Tested subject areas include Survive, First Aid, Equipment Checks, and Defense Measures.

Army History / Army Values (15%)

Army History items cover the history of the U. S. Army and the NCO and include topics such as Courtesy and Customs of the Army, the Volunteer Army, and the End of the Cold War. Also covered in this section are the seven Army Values.

Leadership (13%)

Leadership items cover topics of importance for junior NCOs, including such topics as the Duties and Responsibilities of Officers and NCOs, the Chain of Command and Support Channel, Troop Leading Procedures, Principles of Discipline, and Risk Management.

Training (12%)

The items for the Training section come from topics such as Responsibilities of NCOs in Training, Preparatory Marksmanship Training, and Preparing for and Conducting Drill and Ceremonies.

Study References

Suggested References are listed in the table below. Clicking on a title will link you to the actual manual.

Table 1. Study References

Title	Manual/Publication	Publication Date
<u>The Army Noncommissioned Officer Guide</u>	FM 7-22.7	23 December 2002
<u>Army Leadership: Be, Know, Do</u>	FM 22-100	August 1999
<u>Battle Focused Training</u>	FM 7-1	15 September 2003
<u>Drill and Ceremonies</u>	FM 3-21.5	7 July 2003
<u>Rifle Marksmanship M16A1, M16A2/3, M16A4, and M4 Carbine</u>	FM 3-22.9	24 April 2003
<u>Risk Management</u>	FM 100-14	23 April 1998
<u>Salutes, Honors, and Visits of Courtesy</u>	Army Regulation 600-25	1 September 1983
<u>The Soldier's Guide</u>	FM 7-21.13	February 2004
<u>Soldier's Manual of Common Tasks Skill Level 1</u>	Soldier Training Publication No.21-1-SMCT	February 2003
<u>Soldier's Manual of Common Tasks Skill Levels 2, 3, and 4</u>	Soldier Training Publication No. 21-24-SMCT	February 2003
<u>Training the Force</u>	Field Manual 7-0	22 October 2002

Sample Items

Sample items are provided to allow you to become familiar with the content. You might be wondering what these questions might ask. Figure 1 shows three items that are similar to those you will find on the DCAP.

What is the maximum amount of time a MEDEVAC request should take to transmit?

- ☐ 15 seconds
- ☐ 25 seconds
- ☐ 35 seconds
- ☐ 45 seconds

You are performing maintenance on your M16/M4 rifle. How often should you disassemble and clean the extractor and spring assembly?

- ☐ Only when the parts are dirty or damaged.
- ☐ With every third cleaning of your rifle.
- ☐ Whenever you clean your rifle.
- ☐ Never.

Where is MOST of individual Soldier skill development conducted?

- ☐ Self-development training
- ☐ Unit training
- ☐ Initial entry training (IET) (basic or OSUT)
- ☐ Advanced individual training (AIT)

Figure 1. Sample Items

LeadEx

You will also be taking a related, but different exercise in a separate section – the Leadership Judgment Exercise or LeadEx. The LeadEx is designed to assess supervisory performance dimensions, but it is different from the previously described section because there are no completely right or completely wrong answers in the choices given. *This is not the kind of assessment for which you can study.* So, you do not need to worry about preparing for this section.

Each LeadEx item will present a scenario that junior supervisors or NCOs often face, along with four possible actions that an individual might take. You will be asked to indicate which action you judge as *most* effective, and which action you judge as *least* effective.

Instructions:

Read the situation and the four possible actions that follow. Decide which of the four possible actions is most effective. Then, mark your answer (A, B, C, or D) by clicking the appropriate round, white button across from the word MOST. Next, decide which of the remaining three actions is least effective. Then, mark your answer (A, B, C, or D) by clicking the appropriate round, white button across from the word LEAST. You can select only one action as most effective and only one action as least effective.

The example below illustrates how to respond to these items.

One of your fellow soldiers feels like he doesn't have to pitch in and do the work that you were all told to do. What should you do?

- A. Explain to the soldier that he is part of a team and needs to pull his weight.
- B. Report him to the NCO in charge.
- C. Keep out of it; this is something for the NCO in charge to notice and correct.
- D. Find out why the soldier feels he doesn't need to pitch in.

Suppose that, of the four actions above, you judged action D as the most effective. You would click on the button next to option D across from the word MOST. Next, suppose that, of the remaining three actions, you judged action B as the least effective. You would click on the button next to option B across from the word LEAST.

Most	<input type="radio"/> A	<input type="radio"/> B	<input type="radio"/> C	<input checked="" type="radio"/> D
Least	<input type="radio"/> A	<input checked="" type="radio"/> B	<input type="radio"/> C	<input type="radio"/> D

Figure 2. A Sample LeadEx Item

■ Test-Taking Strategies

1. Begin preparing in advance so that you can study a little at a time. Set goals and pace yourself.
2. Get a good night's rest before taking the pilot test.
3. Read all questions carefully. There are no trick questions, but skipping key words can dramatically change the meaning.
4. Read each response thoroughly before answering. The responses may be only slightly different from each other, but this slight difference may be what makes one clearly correct.
5. Look for words such as *most*, *first*, *last*, *immediately*, *always*, or *never* that can help you choose the best response.
6. You will be able to skip difficult questions to come back to later. Before finishing the test, you should have answered all items. If you do not know the answer to a question – guess.

■ Feedback

What kind of feedback will I receive?

We plan to give you feedback on how you did on the test (percentage correct) and how your performance on the test compared with other Soldiers who took this test. However, these items have never been administered before, and the final format of the test depends on the data received from the pilot test. For example, we may find bad items that have to be thrown out, and that will not be included in anyone's score.

When will I receive my feedback?

The pilot test is expected to run into the fall of 2004. The analyses mentioned above cannot proceed until all the data have been collected. Therefore you may receive your feedback as early as late fall, or as late as the end of this year.

How will I receive my feedback?

Your feedback will be sent to your Army email address. Remember, you, alone, will receive your feedback. It is your decision whether to share your results with your supervisor or buddies.

Contact Us



It is important that you contact us with any questions that you may have now or that arise in the future. This includes any change in your status that might affect your being able to participate as tasked. If leaving a voice mail or sending an email, be sure to indicate you are a "PerformM21 Soldier."

Human Resources Research Organization (HumRRO) Contact:

Karen Moriarty, PhD
Research Scientist
(800) 301-1508 ext. 628
kmoriarty@humrro.org

Army Research Institute for the Behavioral and Social Sciences (ARI) Contact:

Tonia Heffner, PhD
Senior Research Psychologist
Comm: (703) 602-7948
DSN 332-7948
Tonia.heffner@hqda.army.mil

APPENDIX B

Sample Soldier Feedback Report

Dear Soldier:

Recall this past year that you took part in the Demonstration Competency Assessment Program (DCAP) pilot test. Attached are your results. As mentioned during the test administration, your results are confidential. Your feedback report is being provided only to you. No one in your chain of command will receive your individual performance information. So, the decision to share this report is yours to make. We encourage you to carefully review this information. If there is an area in which you did not perform as well as you would have liked, we encourage you to review the relevant manuals in the reference list from the *Test Preparation Guide*.

You may recall that there were two major sections to the DCAP pilot:

- Job Knowledge Questions covering Common Tasks Skill Levels 1 and 2, Army/NCO History, Training, and Leadership
- Leadership Judgment Exercise – each item presented a scenario that junior supervisors or NCOs often face along with four possible actions that may be taken. You selected which action you judged as most effective and which action you judged as least effective.

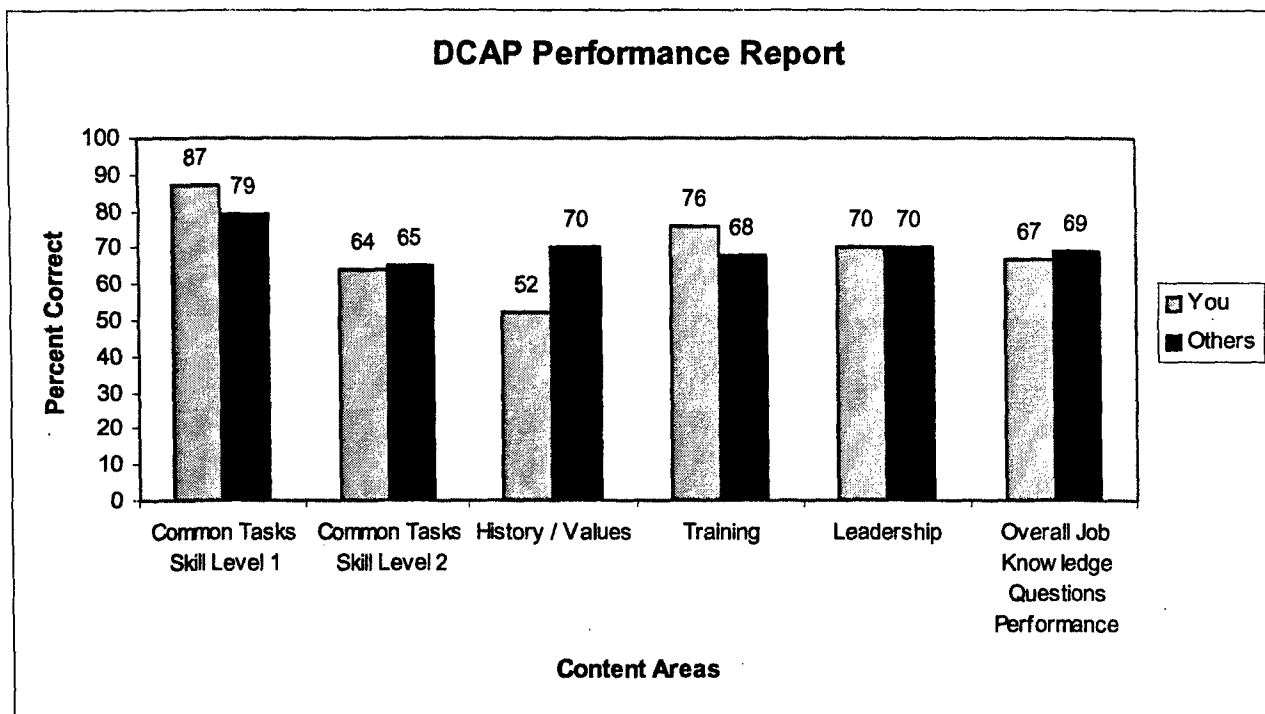
The feedback for each section will be presented separately. The numbers atop the columns represent percentage correct. Your performance is represented by the grey bar, and the pilot sample's performance is represented by the black bar.

If you have any questions about this report, you may contact Karen Moriarty at (800) 301-1508 ext. 628 or karen.moriarty@us.army.mil.

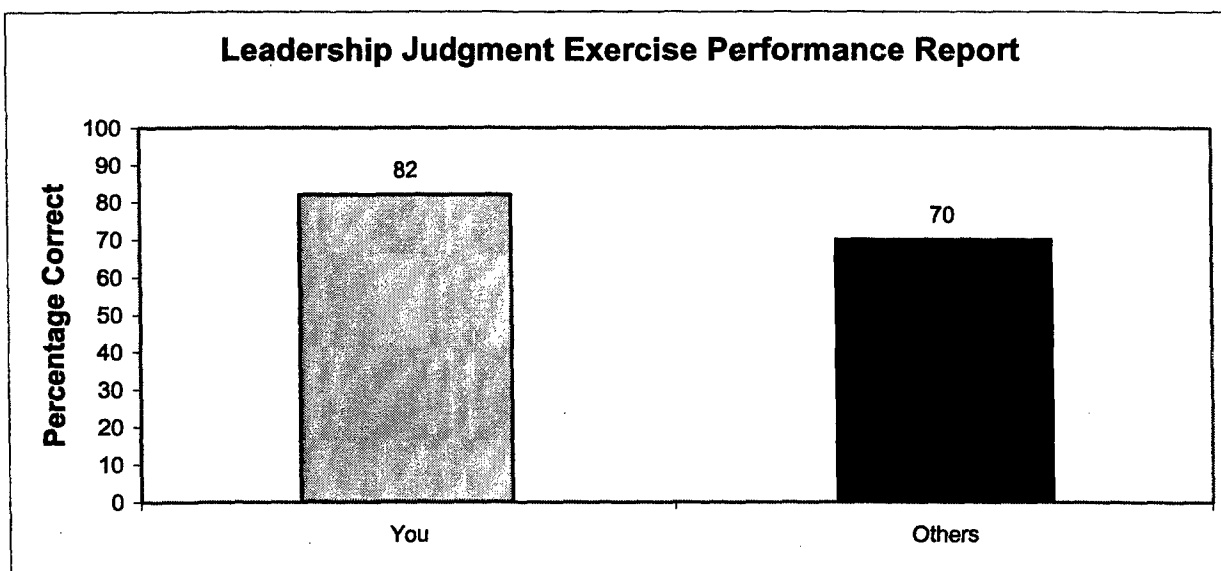
Thank you very much for your participation.

Sincerely,

[Note: The above text will comprise the body of the email. The graphs below will be an attachment.]



Note: Results in this graph do not include performance on the Leadership Judgment Exercise (i.e., the items with the “Most” and “Least” options). For those results see the next graph.



APPENDIX C

Test Item Development Handbook

Prepared by:
The Human Resources Research Organization
November 2004

Job Knowledge Items

Test items may tap into an array of knowledge levels from the concrete to the abstract. The following examples illustrate questions requiring an increasing depth of knowledge:

What does TV stand for?

What is the main function of a TV (television)?

What physical principle is used to display images on a TV (television)?

Your TV (television) is not working properly. What is the most likely cause of the following set of symptoms: ... ?

The levels of knowledge tested in an examination are best described as recall knowledge and applied (procedural) knowledge. These levels are defined below and illustrated through examples.

Knowledge Recall – requires the candidate to recognize, identify, or comprehend a simple fact or term. Recall items generally ask for definitions, common principles, arrangement of steps, or causes.

Example Recall Question

What city is the capital of The United States of America?

- a. Mount Vernon
- b. Philadelphia
- c. Seattle
- d. Washington

Knowledge Application – requires the candidate to understand the reason for each step in an operation and how it relates to other steps. It requires the candidate to be able draw conclusions from a given set of information or apply it to a different situation. Application items are the “why” and “how” questions and generally ask for purposes, causes, interpretations and associations.

Example of Application Question

You are preparing popcorn to eat while you watch the Super Bowl. Suddenly, as you turn on the electric popcorn maker, the lights, the television, and the popcorn maker turn off. What is the first step you take to solving this problem?

- a. Rush to a hotel so you will not miss your game.
- b. Find a flashlight and a book to read.
- c. Check the circuit breakers.
- d. Call the city and complain about your landlord.

Content Validity of Items

Examination Test Content Overview

A test blueprint defines the content, or knowledge domain, that is measured by a test and specifies the percentage of the test that covers a particular domain. It is usually presented as a table listing the knowledge domain and the corresponding number of items on the test.

Validity and Reliability

To ensure reliability and validity of a test, the items should reflect the purpose of the test in content and difficulty, that is:

- Items should measure an understanding or application of the occupational knowledges covered in the test blueprint and course material.
- Items should be realistic and practical; that is, they should be discarded if they are obscure or trivial.
- If an item is so difficult that few test candidates are able to answer it or so easy that most test candidates answer it correctly, the item provides little information in identifying better candidates.

Keep the following questions in mind when you are developing test items:

- How important is the knowledge required to answer this item for the position in question?
- What kind of performance problems might result from lack of the knowledge required for answering this item?

If the knowledge is at least somewhat important and performance problems are moderately damaging, the question most likely meets the criteria listed above.

Traditional (Multiple-choice) Item Anatomy

Stem – The premise of the item. It may be a question, direction, graphic, or incomplete statement.

The STEM should be a complete expression of the problem, missing only the essential information necessary to answer the question. After reading the premise, an informed candidate should have an idea of what the correct answer is before he or she sees the response options.

The stem may ask for a definition, the purpose or cause of an activity, an association or similarity, recognition of an error, or a common principle.

Response options – All the answer choices.

There are always a minimum of three response options (i.e., no true or false questions). Generally there are four response options.

Key – The correct answer.

There should be only one correct answer. It should be clear, grammatically consistent with the stem, and unique from the other response options.

Distractors – The incorrect response options.

Distractors should be unique from one another and plausible enough to distract, or lure the less informed candidate from the correct answer. They should be absolutely incorrect or inferior to the correct answer, according to the source reference. They should be about the same length as the correct answer, similar in form and format, and written to the same level of generality/specificity.

Format of Multiple-choice Items

If the stem ends in an incomplete sentence, then end the stem with a colon and end each response option with a period. For example,

The best topping for vanilla ice cream is:

- a. caramel.
- b. hot fudge.
- c. fruit.
- d. sprinkles.

If the stem ends in a complete sentence, then do not end the response options with a period unless the response options are complete sentences or the options contain more than 3 to 4 words. Capitalize the first word in each option unless it is a special word that is never capitalized. Finally, if one response option should end with a period, then all should. For example,

Who should play quarterback for the Washington Redskins?

- a. Ramsey.
- b. Brunnell.
- c. Johnson.
- d. I do not care.

****Because option d is a complete sentence, it must end in a period. Therefore, all options end in periods.**

When Writing Distractors,

Incorporate:

- < Common misunderstandings of a concept or process
- < Common confusions between similar concepts, processes, or terms
- < Common errors you have seen
- < Using the wrong formula
- < Skipping a step

Use:

- < Familiar, yet incorrect phrases
- < True statements that do not correctly answer the question

Avoid:

- < Fictitious terms. Use common errors.
- < Joke distractors
- < For example, don't use "HMO (Human Misery Organization)" as a distractor.

Submitting to the Item Bank

When submitting items to the item bank, please provide the following information:

- < The blueprint area to which the item belongs
- < The source reference, including the author, edition, and page number
- < An indication of the correct answer
- < The text of the question (the item stem) and any associated graphics
- < Four response options

Guidelines for Writing Multiple-Choice Job Knowledge Items

1. Make sure the response alternatives are grammatically correct and consistent with the stem.
2. Make the question as simply worded and specific as possible. Avoid extraneous material. Avoid writing "teaching" questions.
3. Develop several (3-5) answer choices. Make each response alternative about equal in length.
4. Be sure that only one of the response alternatives is correct.
5. Try to use likely examinee mistakes as distractors (i.e., incorrect options). Always make all distractors plausible choices (not obviously wrong).
6. Make each of the wrong alternatives separate and distinct. Do not offer the wrong alternative in different forms in different alternatives.
7. Response options should have temperament consistency – for example, if the question asks what a professional should *avoid* doing, an incorrect response option should have a *negative* connotation. If the question is asking what the best approach to something is, incorrect response options should have positive connotations.
8. As a general rule, do not construct wrong answers that are direct opposites of the correct answer.
9. Do not use multiple response options (e.g., ANone of the above@, AAll of the above@, Options 1 and 2 are correct").
10. Do not give clues or hints to the correct answer within the item stem.
11. Minimize uses of negatives, both in the stem and the response alternatives -- if you use negatives, underline them to draw attention to that part of the item. Negative questions, for the most part, are distracting and confusing to candidates.
12. Avoid terms like "always", "none", and "all".
13. Avoid using *would do* – use *should do* instead.

Non-Traditional Items

With new technology, we are able to develop "non-traditional" test items. Non-traditional is a term used to describe matching, drag and drop, multiple response, or ranking items. The use of non-traditional items will be limited for paper and pencil assessments. Our research has shown that non-traditional items perform well psychometrically, and test-takers report enjoying the break from multiple-choice items. So, whenever possible, the use of non-traditional items is encouraged.

Much of what is written above for multiple-choice items applies to non-traditional items, but there are a few additional things to keep in mind.

Matching Items

An example of a matching item is –

Match the following states to their capitals.

- | | | |
|-------------------|-----|--------------|
| 1. South Carolina | ___ | a. Frankfort |
| 2. Kentucky | ___ | b. Hartford |
| 3. Massachusetts | ___ | c. Columbia |
| 4. Connecticut | ___ | d. Boston |

It is better to have more options (i.e., capitals in this example) than stimuli (i.e., states in this example). So, a better way to construct this item is –

Match the following states to their capitals.

- | | | |
|-------------------|-----|--------------|
| 1. South Carolina | ___ | a. Raleigh |
| 2. Kentucky | ___ | b. Frankfort |
| 3. Massachusetts | ___ | c. Fargo |
| 4. Connecticut | ___ | d. Hartford |
| | | e. Boston |
| | | f. Columbia |
| | | g. Pierre |

Limit the number of stimuli presented in any one item to 5.

Multiple Response Items

An example of a multiple response item is –

Which of these adults has a blood pressure reading outside the normal range?

Select all that apply.

Adult	BP Reading
Mary	150/90
Bob	120/70
Ellen	100/60
Frank	100/50
Pat	90/50

- a. Mary
- b. Bob
- c. Ellen
- d. Frank
- e. Pat

It is important for these items that you remember to include the “Select all that apply” sentence just below the stem. Also, do not give full credit to someone who simply selects all options.

APPENDIX D

Job Analysis Ratings

Table D1. Generalized Work Activity (GWA) Ratings

	14E		31B		68W	
	(n = 11) Mean	SD	(n = 13) Mean	SD	(n = 21) Mean	SD
Getting Information: Observing, receiving, and otherwise obtaining information from relevant sources.	4.00	0.77	4.54	0.88	4.52	0.813
Contributing to and Supporting Teams: Providing assistance and support to team members and helping the team remain focused on its goals.	4.36	0.67	4.00	1.15	4.41	0.71
Updating and Using Relevant Knowledge: Keeping up-to-date technically and applying new knowledge to your job.	4.45	0.52	3.85	0.99	4.33	0.913
Communicating with Supervisors, Peers, or Subordinates: Providing information to supervisors, coworkers, and subordinates by telephone, in written form, e-mail (or other electronic devices), or in person.	4.09	0.94	4.54	0.66	4.3	1.03
Working with Computers: Using computers, computer systems, or other computer based technology (including hardware and software) to program, write software, set up functions, enter data, or process information.	3.73	1.01	2.38	1.04	4.24	1.04
Operating Vehicles, Mechanized Devices, or Equipment: Running, maneuvering, navigating, or driving vehicles or mechanized equipment, such as forklifts, passenger vehicles, infantry fighting vehicles, tanks, or water craft.	4.00	1.00	3.92	1.26	4.14	1.11
Establishing and Maintaining Interpersonal Relationships: Developing constructive and cooperative working relationships with others, and maintaining them over time.	3.55	0.52	4.38	0.96	4.05	0.944
Evaluating Information to Determine Compliance with Standards: Using relevant information and individual judgment to determine whether events or processes comply with regulations, policies or procedures.	3.82	0.98	3.77	1.01	3.9	1.34
Inspecting Equipment, Structures, or Materials: Inspecting equipment, structures, or materials to identify the cause of errors or other problems or defects.	4.18	0.75	3.85	1.14	3.86	1.11
Assisting and Caring for Others: Providing personal assistance, medical attention, emotional support, or other personal care to others such as coworkers, indigenous personnel, patients, or clients.	3.91	1.30	3.85	1.07	3.75	1.25
Controlling Machines and Processes: Using either control mechanisms or direct physical activity to operate machines or processes (not including computers or vehicles).	3.36	0.92	2.62	1.39	3.67	1.11
Identifying Objects, Actions, and Events: Identifying information by categorizing, estimating, recognizing differences or similarities, and detecting changes in circumstances or events.	3.64	0.81	4.46	0.97	3.67	1.2

Table D1. (Continued)

	14E		31B		68W	
	(n = 11)		(n = 13)		(n = 21)	
	Mean	SD	Mean	SD	Mean	SD
Making Decisions and Solving Problems: Analyzing information and evaluating results to choose the best solution and solve problems in situations where judgment is required (i.e., when the problem cannot be straightforwardly resolved by applying a specific knowledge[s] or skill[s]).	4.18	0.60	3.62	0.87	3.67	1.29
Interpreting the Meaning of Information for Others: Translating or explaining what information means and how it can be used.	3.09	1.22	3.23	0.83	3.65	1.04
Organizing, Planning, and Prioritizing Work: Developing specific goals and plans to prioritize, organize, and accomplish your work.	3.27	0.79	2.69	1.11	3.53	1.25
Repairing and Maintaining Mechanical Equipment: Servicing, repairing, adjusting, and testing machines, devices, moving parts, and equipment that operate primarily on the basis of mechanical (not electronic) principles.	4.09	1.04	3.31	1.11	3.53	1.29
Monitoring Processes, Materials, or Surroundings: Monitoring and reviewing information from materials, events, or the environment, to detect or assess problems.	3.64	0.92	4.38	0.87	3.52	0.98
Performing Administrative Activities: Performing day-to-day administrative tasks such as maintaining information files and processing paperwork.	2.91	0.94	2.85	0.99	3.47	0.8
Judging the Qualities of Objects, Services, or People: Assessing the value, importance, or quality of things or people.	3.27	1.19	3.77	1.17	3.38	1.07
Estimating the Quantifiable Characteristics of Products, Events, or Information: Estimating sizes, distances, and quantities; or determining time, resources, or materials needed to perform a work activity.	3.18	0.40	3.54	1.05	3.38	1.16
Analyzing Data or Information: Identifying the underlying principles, reasons, or facts of information by breaking down information or data into separate parts.	3.18	0.40	2.85	0.99	3.33	1.2
Communicating with People Outside the Organization: Communicating with people outside the organization, representing the organization to the public, other parts of the military or government, and/or indigenous personnel. The information can be exchanged in person, in writing, or by telephone or e-mail (or other electronic devices).	2.64	0.81	3.85	1.41	3.25	1.25
Processing Information: Compiling, coding, categorizing, calculating, tabulating, auditing, or verifying information or data.	3.55	0.69	3.15	1.07	3.14	1.11
Thinking Creatively: Developing, designing, or creating new applications, ideas, relationships, systems, or products, including artistic contributions.	2.55	0.82	2.54	0.88	3.05	1.07
Influencing Others: Convincing others to change their minds or actions.	3.36	1.29	3.08	1.50	2.7	1.3

Table D1. (Continued)

	14E		31B		68W	
	(n = 11)		(n = 13)		(n = 21)	
	Mean	SD	Mean	SD	Mean	SD
Monitoring and Controlling Resources: Monitoring and controlling the expenditure of resources.	2.91	0.94	1.92	1.04	2.47	1.23
Drafting, Laying Out, and Specifying Technical Devices, Parts, and Equipment: Providing documentation, detailed instructions, drawings, or specifications to tell others about how devices, parts, equipment, or structures are to be fabricated, constructed, assembled, modified, maintained, or used.	2.45	1.04	2.00	1.08	2.43	1.03
Scheduling Activities: Scheduling events, programs, and activities.	2.55	0.69	1.69	1.11	2.38	1.2
Repairing and Maintaining Electronic Equipment			2.46	1.33		

Note. Raters used a 5-point importance scale (1 = not important, 5 = extremely important). The last GWA listed was only shown to the 31B sample.

Table D2. Cognitive Complexity Ratings

	14E						31B						68W					
	(n = 11)			(n = 24)			(n = 7)			(n = 7)			(n = 7)			(n = 7)		
	Importance M	SD	Frequency M	Importance M	SD	Frequency M	Importance M	SD	Frequency M	Importance M	SD	Frequency M	Importance M	SD	Frequency M	Importance M	SD	Frequency M
Judgment/Problem Solving: Situations where the Soldier needs to exercise judgment regarding problems not easily addressed by regulations, rules, manuals, SOPs, or Rules of Engagement.	4.50	0.71	4.30	0.82	0.88	3.99	0.88	0.93	4.86	.38	4.41	1.46						
Information Intensity: Situations involving complex decision-making and/or processing a lot of information at one time.	4.80	0.42	4.80	0.42	0.75	4.19	0.75	0.80	4.71	.76	3.57	1.40						
Systems Thinking: Situations where Soldiers must understand a "system" with a number of interrelated elements that affect each other.	5.00	0.00	5.00	0.00	0.97	3.25	0.97	1.17	4.29	1.11	3.71	1.25						

Note. Raters used a 5-point importance scale (1 = not important, 5 = extremely important) and a 5-point frequency scale (1 = never, 5 = everyday).

Table D3. Work Context Ratings

	14E						31B						68W					
	(n = 11)			(n = 24)			(n = 10)			(n = 10)			(n = 10)			(n = 10)		
	Importance M	SD	Frequency M	Importance M	SD	Frequency M	Importance M	SD	Frequency M	Importance M	SD	Frequency M	Importance M	SD	Frequency M	Importance M	SD	Frequency M
Methods of Getting Information																		
Face-to-Face	4.80	0.42	5.00	4.70	0.48	0.00	4.70	0.48	0.00	4.70	0.48	0.00	4.70	0.48	0.00	4.70	0.48	0.00
Radio/Telephone or Other Audio Devices	4.90	0.32	4.70	4.69	0.48	0.48	4.69	0.48	0.48	4.69	0.48	0.48	4.69	0.48	0.48	4.69	0.48	0.48
Manuals, Memos, or Other Non-computerized Text	4.50	0.71	4.40	3.84	0.68	0.70	3.84	0.68	0.70	3.84	0.68	0.70	3.84	0.68	0.70	3.84	0.68	0.70
Computer/Visual Display	4.90	0.32	4.90	3.17	0.93	0.32	3.17	0.93	0.32	3.17	0.93	0.32	3.17	0.93	0.32	3.17	0.93	0.32
Extreme Conditions/Hazards																		
Vigilance	4.00	1.25	4.50	4.12	0.92	0.53	4.12	0.92	0.53	4.12	0.92	0.53	4.12	0.92	0.53	4.12	0.92	0.53
Attention to Detail	4.70	0.48	4.90	4.83	0.37	0.32	4.83	0.37	0.32	4.83	0.37	0.32	4.83	0.37	0.32	4.83	0.37	0.32
Time Pressure	5.00	0.00	5.00	4.61	0.48	0.00	4.61	0.48	0.00	4.61	0.48	0.00	4.61	0.48	0.00	4.61	0.48	0.00
Interpersonal Conflict	5.00	0.00	4.70	4.41	0.65	0.48	4.41	0.65	0.48	4.41	0.65	0.48	4.41	0.65	0.48	4.41	0.65	0.48
Social Interaction	3.90	1.10	4.30	4.68	0.40	0.48	4.68	0.40	0.48	4.68	0.40	0.48	4.68	0.40	0.48	4.68	0.40	0.48
Distractions and Interruptions	4.90	0.32	4.90	4.68	0.34	0.32	4.68	0.34	0.32	4.68	0.34	0.32	4.68	0.34	0.32	4.68	0.34	0.32
	4.20	0.92	4.70	4.14	0.78	0.48	4.14	0.78	0.48	4.14	0.78	0.48	4.14	0.78	0.48	4.14	0.78	0.48

Note. Raters used a 5-point importance scale (1 = not important, 5 = extremely important) and a 5-point frequency scale (1 = never, 5 = everyday).

Table D4. Knowledge and Skill Ratings

	14E						31B						68W					
	(n = 11)			(n = 24)			(n = 11)			(n = 11)			(n = 11)			(n = 11)		
	Importance M	Frequency SD	Importance M	Frequency SD	Importance M	Frequency SD	Importance M	Frequency SD	Importance M	Frequency SD	Importance M	Frequency SD	Importance M	Frequency SD	Importance M	Frequency SD	Importance M	Frequency SD
Declarative Knowledge	5.00	0.00	4.82	0.60	4.38	0.54	4.64	0.65	4.45	.82	4.27	1.01						
Procedural Knowledge & Skills																		
Reading Skill	4.36	0.50	4.64	0.50	4.36	0.32	4.29	0.52	4.78	.44	5.00	.00						
Physical Skill	4.36	0.81	4.09	0.83	4.46	0.45	4.58	0.67	4.73	.47	4.36	.67						
Psychomotor Skill	3.09	1.14	3.00	1.10	4.39	0.73	3.74	0.81	4.73	.47	4.18	.60						
Speaking Skill	4.45	0.82	4.64	0.67	4.77	0.26	4.86	0.23	4.67	.50	4.89	.33						
Writing Skill	3.09	0.70	3.45	0.93	4.24	0.52	4.19	0.61	4.67	.71	4.44	.88						
Listening Skill	4.36	0.81	4.73	0.47	4.68	0.25	4.73	0.40	4.55	.69	4.91	.30						
Interpersonal Skill	4.55	0.52	4.91	0.30	5.00	0.00	4.85	0.38	4.18	.87	4.27	.79						
Using Information Resources	4.27	0.65	4.36	0.50	3.84	0.77	3.74	0.75	4.00	.87	3.78	.83						

Note. Raters used a 5-point importance scale (1 = not important, 5 = extremely important) and a 5-point frequency scale (1 = never, 5 = everyday).